Barts
Cancer Institute
Queen Mary, University of London

Barts and The London
School of Medicine and Dentistry

The Genetics of Familial Leukaemia and Myelodysplasia

Ahad Fahad H Al Seraihi

A thesis submitted for the Degree of Doctor of Philosophy (PhD)

at Queen Mary University of London

January 2019

Centre for Haemato-Oncology

Barts Cancer Institute

Charterhouse Square

London, UK

EC1M 6BQ

Statement of Originality

I, Ahad Fahad H Al Seraihi, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

30th January 2019

Details of Collaborations:

- Targeted deep sequencing detailed in **Chapter 3** was carried out at King's College Hospital NHS Foundation Trust, London, UK, in the laboratory for Molecular Haemato-Oncology led by Dr Nicholas Lea and bioinformatics analysis was performed by Dr Steven Best.

- RNA sequencing in **Chapter 3** was performed at Oxford Genomics, UK, and bioinformatics analyses were carried out by Dr Jun (Alex) Wang and Dr Ai Nagano at the Barts Cancer Institute, Queen Mary University of London, London, UK.

- Final steps of Sanger sequencing (capillary electrophoresis) were carried out at GATC Biotech (Eurofins) services based in Germany.

- Samples from 3 *GATA2*-mutated families with reduced penetrance were kindly provided by Dr Marcin Wlodarski at the Department of Pediatric Hematology and Oncology, University Children's Hospital Freiburg, Freiburg, Germany.

- **Section 1.1** of the Introduction (**Chapter 1**) was reproduced from a book chapter entitled: "Genomics and Diagnostics in Acute Myeloid Leukaemia" published by *The Royal Society of Chemistry* in which I was first author (see details below).

- Much of the work described in the Results sections (**Chapter 3** and **Chapter 4**) was published in the journal *Leukemia* – Springer Nature (**Appendix 1**) in which I was first author jointly with Dr Ana Rio-Machin (see details below). I performed and analysed all the experiments while Ana contributed to the study design and analysis.

Details of Publications:

- **Al Seraihi AF**, Rio-Machin A, Tawana K, Charrot S, Di Bella D, Bödör C, Butler T, Farren T, Grantham M and Fitzgibbon J. Genomics and Diagnostics in Acute Myeloid Leukaemia. Book Chapter in: Whitehouse D and Rapley R (eds) Genomics and Clinical Diagnostics. The Royal Society of Chemistry, London. (*Book is scheduled for release on 29 January 2019*).

- **Al Seraihi AF**, Rio-Machin A, Tawana K, Bödör C, Wang J, Nagano A, Heward JA, Iqbal S, Best S, Lea N, McLornan D, Wlodarski MW, Kozyra E, Niemeyer CM, Scott H, Ellison A, Tummala H, Cardoso SR, Vulliamy T, Dokal I, Smith M, Cavenagh J and Fitzgibbon J. *GATA2* monoallelic expression underlies reduced penetrance in inherited *GATA2*-mutated MDS/AML. *Leukemia.* 2018. 10.1038/s41375-018-0134-9.

- Tummala H, Dokal AD, Walne AJ, Ellison A, Cardoso S, Kirwan M, Browne I, Sidhu JK, rajeeve V, Rio-Machin A, **Al Seraihi A**, Duncombe AS, Jenner M, Smith OP, Enright H, Norton A, Aksu T, Ozbek NY, Pontikos N, Cutillas PR, Dokal I and Vulliamy T. Genome instability is a consequence of transcription deficiency in bone marrow failure patients harboring biallelic *ERCC6L2* variants. *PNAS.* 2018.

- Cardoso SR, Ellison A, Walne A, Cassiman D, Raghavan M, Kishore B, Ancliff P, Rodriguez-Vigil C, Dobbels B, Rio-Machin A, **Al Seraihi A**, Pontikos N, Tummala H, Vulliamy T and Dokal I. Myelodysplasia and liver disease extend the spectrum of *RTEL1* related telomeropathies. *Heamatologica.* 2017; 8:e293-e296.

- Tawana K, Wang J, Király PA, Kállay K, Benyó G, Zombori M, Csomor J, **Al Seraihi A**, Rio-Machin A, Matolcsy A, Chelala C, Cavenagh J, Fitzgibbon J and Bödör C. Recurrent somatic JAK-STAT pathway variants within a novel *RUNX1*-mutated pedigree. *Eur J Hum Genet.* 2017; 8:1020-1024.

- Cardoso SR, Ryan G, Walne AJ, Ellison A, Lowe R, Tummala H, Rio-Machin A, Collopy L, **Al Seraihi A**, Wallis Y, Page P, Akiki S, Fitzgibbon J, Vulliamy T and Dokal I. Germline heterozygous *DDX41* variants in a subset of familial myelodysplasia and acute myeloid leukemia. *Leukemia*. 2016; 10:2083-2086.

- Okosun J, Wolfson RL, Wang J, Araf S, Wilkins L, Castellano BM, Escudero-Ibarz L, **Al Seraihi AF**, Richter J, Bernhart SH, Efeyan A, Iqbal S, Matthews J, Clear A, Guerra-Assunção JA, Bödör C, Quentmeier H, Mansbridge C, Johnson P, Davies A, Strefford JC, Packham G, Barrans S, Jack A, Du MQ, Calaminici M, Lister TA, Auer R, Montoto S, Gribben JG, Siebert R, Chelala C, Zoncu R, Sabatini DM and Fitzgibbon J. Recurrent mTORC1-activating *RRAGC* mutations in follicular lymphoma. *Nature Genetics.* 2016; 48:183–188.

Manuscripts Pending Review or in Preparation:

- Heward JA, Konali L, D'Avola A-L, Yeomans A, Rahim T, **Al Seraihi AF**, Wang J, Korfi K, Araf S, Iqbal S, Calaminici M, Clear A, Okosun J, Johnson P, Gribben J, Packham G, Fitzgibbon J. KDM5 inhibition is a novel therapeutic strategy for the treatment of *KMT2D* mutant lymphomas. (*Submitted in December 2018 – in review*).

Conference Presentations

Oral Communications:

- **Al Seraihi A.** *Reduced Penetrance in GATA2-mutated Familial Leukaemia.* Barts Cancer Institute Annual Postgraduate Day. London, UK. June 2018.

- **Al Seraihi A**. *Variable Penetrance in GATA2 Deficiency*. 3rd Saudi Hematology Research Day. Saudi Hematology Annual Congress. Riyadh, Saudi Arabia. March 2018.

- **Al Seraihi A**. *The Genetics of Familial MDS/AML: Challenges and Insights.* Saudi Hematology Year Review 2018. Saudi Society of Blood & Marrow Transplantation. Riyadh, Saudi Arabia. February 2018.

- **Al Seraihi A**. *Germline GATA2 P.T354M mutations: why so shy?* Barts Cancer Institute Seminar Series. London, UK. July 2017.

- **Al Seraihi A**. *Variable Penetrance in GATA2 Deficiency.* Friday Scientific Workshop on Inherited Hematopoietic Malignancies. 58th American Society of Hematology (ASH) Annual Meeting and Exposition, San Diego, CA, USA. December 2016.

Posters:

- Rio-Machin A, Cardoso S, Ellison A, Tawana K, Wang J, Chelala C, Plagnol V, Pontikos N, Page P, Wallis Y, Ryan G, **Al Seraihi A**, Walne A, Tummala H, Fitzgibbon J, Dokal I and Vulliamy T. *Identifying new disease genes in familial myeolodisplasia/acute myeloid leukemia.* 23rd Congress of the European Haematology Association (EHA), Stockholm, Sweden. June 2018.

- **Al Seraihi A**, Rio-Machin A, Tawana K, Bödör C, Heward J, Iqbal S, Ellison A, Tummala H, Cardoso SR, Cavenagh J, Vulliamy T, Dokal I and Fitzgibbon J. *GATA2 Monoallelic Expression Underlies Reduced Penetrance in GATA2 Deficiency*. 4th International Conference on Acute Myeloid Leukemia "Molecular and Translational": Advances in Biology and Treatment, European Society of Hematology (ESH). Estoril, Portugal. October 2017.

- **Al Seraihi A**, Rio-Machin A, Wang J and Fitzgibbon J. *GATA2 monoallelic expression underlies reduced penetrance in inherited GATA2-mutated MDS/AML. Barts Cancer Institute Annual Postgraduate Day*. London, UK. June 2017.

- **Al Seraihi A**, Rio-Machin A, Tawana K, Bödör C, Araf S, Heward J, Smith M, Iqbal S, Best S, Lea N, McLornan D, Ellison A, Tummala H, Cardoso SR, Cavenagh J, Vulliamy T, Dokal I and Fitzgibbon J. *Variable Penetrance is Linked with Monoallelic Gene Expression in Inherited GATA2-mutated MDS/AML.* 58th American Society of Hematology (ASH) Annual Meeting and Exposition, San Diego, CA, USA. December 2016.

- Rio-Machin A, Cardoso SR, Tawana K, Wang J, Chelala C, Plagnol V, Wallis Y, Ryan G, Ellison A, **Al Seraihi A**, Walne A, Tummala H, Fitzgibbon J, Dokal I and Vulliamy T. *Whole Exome sequencing reveals novel candidate genes in familial MDS/AML*. 21st Congress of the European Haematology Association (EHA), Copenhagen, Denmark. June 2016.

- Cardoso SR, Ryan G, Walne A, Tummala H, Rio-Machin A, Ellison A, Collopy L, **Al Seraihi A**, Wallis Y, Page P, Akiki S, Fitzgibbon J, Vulliamy T and Dokal I. *Germline heterozygous DDX41 variants account for a subset of familial myelodysplasia and acute myeloid leukaemia.* 36th World Congress of the International Society of Hematology and the British Society of Haematology (ISH/BSH). Glasgow, Scotland. April 2016.

Abstract

Background & Aim: While the majority of myelodysplasia and acute myeloid leukaemia (MDS/AML) cases are sporadic, rare familial predisposition syndromes have been delineated and are regarded a separate disease entity in the 2016 WHO classification system. Germline mutations in 14 disease genes have been uncovered thus far, with *GATA2* representing one of the key transcriptional regulators commonly mutated in inherited leukaemias. The rarity of these familial cases opens the door to fundamental questions in biology, one of which is the phenomenon of reduced penetrance posing a clinical challenge particularly when identifying "silent" mutation carriers for genetic screening and exclusion as potential stem cell transplant donors. We have noted that this is indeed a feature within certain *GATA2*-mutated families, especially those carrying germline missense mutations such as (p.Thr354Met). In our example, two first-degree cousins developed MDS/AML with monosomy 7 while a third cousin presented with significant monocytopenia and neutropenia. This contrasted with the parental generation mutation carriers who all remain symptom free into their mid-late 60s. This thesis therefore sets out to investigate the molecular mechanisms underlying the reduced penetrance and clinical heterogeneity observed within a *GATA2*-mutated family with a view of identifying molecular features that distinguish between these two groups of mutation carriers.

Results: Deep targeted sequencing of 33 genes frequently mutated in MDS/AML revealed acquisition of somatic *ASXL1* mutation (p.Gly646TrpfsTer12) in all affected cousins with no mutations detected in asymptomatic family members. It was noteworthy that the variant allele frequency was lower (12%) in the third cousin symptomatic carrier and remained stable (range 12-6%) over a 6-year monitoring period. Total *GATA2* expression was lower in the symptomatic compared with asymptomatic carriers as assessed by RT-qPCR and remarkably this was associated with monoallelic expression favouring the mutant *GATA2* allele with loss of the wild-type (WT) allele expression. Temporal analysis of the symptomatic carrier over a 6-year disease period demonstrated a reactivation of the WT allele expression 3 years later, coinciding with a persistent improvement in haematological parameters. We believe these allele-specific changes in *GATA2* expression are driven by dynamic epigenetic reprogramming that include changes in DNA methylation and chromatin

mark deposition. Using a SNP (rs1806462 [**C**/**A**]) that generates/removes a CpG dinucleotide within *GATA2* promoter region, we first assessed allele-specific differences in DNA methylation by bisulphite sequencing. This demonstrated a significant increase in promoter methylation in the WT allele that returned to normal levels at later time-points. We then assessed allele-specific deposition of H3K4me3 and H3K27me3 chromatin marks by chromatin immunoprecipitation (ChIP). Sanger sequencing revealed a significant enrichment in the deposition of H3K4me3 activating mark on the mutant allele at diagnosis that was reversed at later follow-up, correlating with reactivation of the WT allele expression.

Conclusion: Reduced penetrance is a feature of many families with inherited forms of MDS/AML which may be governed by the acquisition of additional co-operating mutations (e.g. *ASXL1*). In this thesis, however, we show that changes in the WT:mutant *GATA2* allele expression ratio as a result of local and allele-specific changes in DNA methylation and chromatin mark deposition may also influence the penetrance of the germline mutation, adding another layer of complexity to the (epi)genetic basis of familial MDS/AML.

Keywords: Familial MDS/AML, GATA2, reduced penetrance, epigenetics, gene expression.

Dedication

- This thesis is dedicated to the *GATA2*-mutated family members who have greatly contributed to this study by kindly donating their samples for research and all the families affected by inherited forms of leukaemia and myelodysplasia, in particular;
- To Leah, a beautiful brave girl who was diagnosed with myelodysplasia and *GATA2* deficiency at 15 years old and sadly left us way too soon, may she rest in peace.

Acknowledgements

"All scientific careers are voyages of discovery; the journey more important than the uncertain destination. And laboratories are Beagles for their occupants."
– Mel Greaves

This thesis is a result of my work as a PhD student within the Centre for Haemato-Oncology at the Barts Cancer Institute, London (2015-2018) and I am grateful for the handful of amazing people who have helped make this journey both possible and enjoyable along the way:

First and foremost, I would like to express my sincerest appreciation to my primary supervisor, Professor Jude Fitzgibbon, for taking a chance on me as a Master's student and giving me the opportunity to pursue a PhD in his laboratory. Not only did he provide a fantastic lab training and tremendous academic mentorship, he also instilled in me valuable lessons – and clever analogies – on what makes a good scientist/communicator and the importance of asking the right question and telling a story effectively and for that I thank him wholeheartedly; I learnt from the best! My immense gratitude also goes to my secondary supervisor, Professor Inderjeet Dokal, for all his support and guidance throughout my PhD.

I would like to thank everybody in the Centre for Haemato-Oncology, in particular Dr Ana Rio-Machin, for her insightful scientific advice and brilliant suggestions that supported this thesis immensely. Thank you, Ana, for being the best role model one can be, always inspiring me with your energy and love for science and for your friendship, we make a good team. Special thanks also go to Dr Jessica Okosun and Dr Kiran Tawana for holding my hand in the lab at the beginning. It was through my MSc project investigating patterns of clonal evolution in Follicular Lymphoma in Spring 2014, under Jude's and Jessica's guidance, that I decided to continue learning about the molecular pathogenesis of blood cancers for the ensuing 3-4 years and quite possibly, for the rest of my academic career. I am also appreciative of Kiran's help throughout this PhD, whose clinical knowledge of familial leukaemia is impeccable.

To past and current members of our group: Csaba Bödör, Lola Koniali, James Heward, Shamzah Araf, Koorosh Korfi, Joshua Dawkins, Tahrima Rahim, Doriana Di Bella, Emil Kumar, Findlay Copley and Karina Close; thank you all for your camaraderie and support. It has been a pleasure working alongside you and I will always cherish the memories we shared together at Barts.

I am also grateful to Dr Jun (Alex) Wang for his bioinformatics expertise and to Ms Sameena Iqbal and her team for guarding patients' samples in the Tissue Bank and providing them to us ever so swiftly.

The uniqueness of familial leukaemia provided the opportunity to collaborate with select groups both nationally and worldwide; special thanks to the inherited bone marrow failure group at the Blizard Institute - QMUL, led by Professors Inderjeet Dokal and Tom Vulliamy and their team: Hemanth Tummala, Alicia Ellison and Shirleny Cardoso, for their valuable insights and fruitful meeting discussions.

Finally, this PhD would not have been possible without the generous support from the Saudi Arabian Ministry of Higher Education (MoHE) and the Saudi Arabian Cultural Bureau (SACB) in London.

Last, but by no means least, sincerest thanks to my family: Mum, Dad and my brother Faisal who, despite being thousands of miles away (3078 to be exact), continue to shower me with unconditional love and encouragement, every day – شكراً على كل شيء...أحبكم

***

Table of Contents

List of Tables

Chapter 1. Introduction

Chapter 2. Materials & Methods

Chapter 3. Results 1

List of Figures

Chapter 1. Introduction

Chapter 2. Materials & Methods

Chapter 3. Results 1

Chapter 4. Results 2

Chapter 5. Overall Discussion

List of Abbreviations

ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
APL	Acute Promyelocytic Leukaemia
ASD	Autism Spectrum Disorder
ASE	Allele-Specific Expression
ATRA	all-*trans* Retinoic Acid
aUPD	Acquired Uniparental Disomy
BM	Bone Marrow
BMT	Bone Marrow Transplant
BSP	Bisulphite-Specific PCR
CBC	Complete Blood Count
CBF	Core Binding Factor
cDNA	Complementary Deoxyribonucleic Acid
CEBPA	CCAAT/enhancer-binding protein alpha
CFU	Colony Forming Unit
CGH	Comparative Genomic Hybridisation
CHIP	Clonal Haematopoiesis of Indeterminate Potential
ChIP	Chromatin Immunoprecipitation
CMML	Chronic Myelomonocytic Leukaemia
CNA	Copy Number Alteration
COSMIC	Catalogue of Somatic Mutations in Cancer
CpG	Cytosine-phosphate-Guanine
CPM	Count Per Million
CQN	Conditional Quantile Normalisation
CR	Complete Remission
CR1	First Complete Remission
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRUK	Cancer Research United Kingdom
C_t	Cycle Threshold
DA	Daunorubicin
DC	Dyskeratosis Congenita
DCML	Dendritic Cell, Monocyte, B- and natural killer-Lymphoid
DDL	Donor-Derived Leukaemia
DDX41	DEAD/H-box Polypeptide 41
DE	Differential Expression/ Differentially Expressed
Del	Deletion
DMSO	Dimethyl Sulfoxide
DMF	Dimethylformamide
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
EB	Elution Buffer
ELN	European Leukemia Net
EMSA	Electrophoretic Mobility Shift Assay

ES	Embryonic Stem
ESA	Erythroid Stimulating Agents
ESC	Embryonic Stem Cells
ExAC	Exome Aggregation Consortium
FA	Fanconi Anaemia
FAB	French-American-British
FC	Fold Change
FDR	False Discovery Rate
FISH	Fluorescence *in situ* Hybridisation
FPD/AML	Familial Platelet Disorder/Acute Myeloid Leukaemia
GATA2	GATA Binding Factor 2
G-CSF	Granulocyte-Colony Stimulating Factor
GEO	Gene Expression Omnibus
GLM	Generalised Linear Model
GoF	Gain-of-Function
GSEA	Gene Set Enrichment Analysis
GvHD	Graft versus Host Disease
GWAS	Genome Wide Association Studies
HAT	Histone Acetyl Transferase
HD	Hodgkin Disease
HDAC	Histone Deacetylase
HLA	Human Leukocyte Antigen
HSC	Haematopoietic Stem Cell
HSCT	Haematopoietic Stem Cell Transplantation
IBMFS	Inherited Bone Marrow Failure Syndrome
Indel	Insertion/Deletion
ICGC	International Cancer Genome Consortium
IPSS-R	International Prognostic Scoring System - Revised
IVF	*in vitro* Fertilisation
KDM	Lysine Demethylase
KMT	Lysine Methyltransferase
LAIP	Leukaemia-associated Immunophenotype
LAR II	Luciferase Assay Reagent II
LB	Luria Bertani
LD	Linkage Disequilibrium
LIC	Leukaemia initiating cell
LoF	Loss-of-Function
LOH	Loss of Heterozygosity
LLS	Leukaemia and Lymphoma Society
bZIP	Leucine Zipper Region
MAF	Minor Allele Frequency
MDS	Myelodysplastic Syndrome
MeDIP	Methylated DNA Immunoprecipitation
MFC	Multiparameter Flow Cytometry
MIRAGE	MDS, infection, restriction of growth, adrenal hypoplasia, genital phenotypes and enteropathy

MLL	Mixed-Lineage Leukaemia
MLPA	Multiplex Ligation Dependent Probe Amplification
MM	Multiple Myeloma
MonoMac	Monocytopenia and Mycobacterium Infection
MPN	Myeloproliferative Neoplasms
MRD	Minimal Residual Disease
mRNA	Messenger Riboxynucleic Acid
MSP	Methylation-Specific PCR
5mC	5-methylcytosine
5hmc	5-hydroxymethylcytosine
NEB	New England BioLabs
NES	Normalised Enrichment Score
NGS	Next Generation Sequencing
NHL	Non-Hodgkin's Lymphoma
NK	Normal Karyotype
NK-AML	Normal Karyotype Acute Myeloid Leukaemia
NMD	Nonsense-mediated Decay
NTC	Non-template Control
OS	Overall Survival
PB	Peripheral Blood
PBS	Phosphate-Buffered Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PGD	Preimplantation Genetic Diagnosis
pH	Power/Potential of Hydrogen
PRC1/2	Polycomb Repressive Complex1/2
PTD	Partial Tandem Duplication
PTM	Post-Translational Modification
RA	Refractory Anaemia
RAEB-1/2	Refractory Anaemia with Excess Blasts-1/2
RHD	Runt Homology Domain
RIN	RNA Integrity Number
RNA	Riboxynucleic Acid
RT-qPCR	Real-Time Quantitative Polymerase Chain Reaction
SAP	Shrimp Alkaline Phosphatase
SD	Standard Deviation
SDS	Shwachman-Diamond Syndrome
SEM	Standard Error of the Mean
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
T_A	Annealing Temperature
TAD	Transactivation Domain
TBD	Telomere Biology Disorder
TBE	Tris-Borate-EDTA
TCGA	The Cancer Genome Atlas
TERC	Telomerase RNA Component

TERT	Telomerase Reverse Transcriptase
TF	Transcription Factor
T_m	Melting Temperature
TRM	Transplant Related Mortality
TSS	Transcriptional Start Site
TSCA	True SeqCustom Amplicon
UCSC	University of California Santa Cruz
UK	United Kingdom
UTR	Untranslated Region
UV	Ultra Violet
VAF	Variant Allele Frequency
VUS	Variant of Uncertain Significance
WCC	White Blood Cell Count
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organisation
WT	Wild-type
Yrs.	Years
ZF1	First Zinc Finger
ZF2	Second Zinc Finger

Chapter 1. Introduction

1. Introduction

"I start with the premise that all human disease is genetic"

– Paul Berg

Myelodysplastic syndrome (MDS) and acute myeloid leukaemia (AML) are clinically diverse and genetically heterogeneous haematopoietic stem cell malignancies characterised by defective haematopoiesis and premature mortality in many patients (Grimwade et al., 2016). Whilst the majority of MDS/AML cases occur *de-novo* or secondary to chemo- or radio-therapy (**sporadic MDS/AML**), there are rare occurrences of familial cases (<5% in adults – 4-13% in paediatrics) where disease manifestations run in at least two members within the same family (**familial MDS/AML**) (Akpan et al., 2018). These cases represent a high-risk group of patients who require a high-index of clinical suspicion and unique follow-up for comprehensive management and care.

The relative rarity of familial MDS/AML cases probably reflects that many patients in this group are not recognised by the current diagnostic algorithms. This is hindered by the lack of accessible family history and marked variations in disease latency, phenotype and outcome between and within families. This is also accompanied by a paucity of comprehensive diagnostic guidelines and certified genetic testing to meet the clinical need of this rare patient population, ultimately leading to deficiencies in patient care, particularly when considering unsuitable related donors for allogeneic haematopoietic stem cell transplantation (HSCT) with several reports describing donor-derived episodes of AML (Xiao et al., 2011, Berger et al., 2017, Galera et al., 2018). The overarching goal of the research community in this field therefore is to establish a multi-disciplinary roadmap combining efforts from clinicians, genetic counsellors and scientists working together to improve diagnosis, screening and management of these patients.

Fortunately, we are making significant progress in delineating the germline mutational repertoire of these inherited leukaemias; thanks to advances in next generation sequencing (NGS) technologies, germline mutations in approximately a dozen genes, primarily in the myeloid transcription factors (*RUNX1*, *CEBPA*, *GATA2* and *ETV6*) have been implicated thus far, corresponding to ~60% of familial cases (Nickels et al., 2013, Godley, 2014, Babushok et al., 2016, Drazer et al., 2016, Wlodarski and Niemeyer, 2017). These technologies are also offering an unprecedented opportunity to identify novel disease-causing alleles in cases with unknown aetiology and decipher the intra- and inter-familial disease heterogeneity and evolution. Despite these advances, however, assigning pathogenicity to new variants is further complicated by the scarcity of familial cases and the enormous genetic and phenotypic diversity and our knowledge of the functional impact of some of these newly-discovered variants is still very much in its infancy.

Another challenge to contend with is the **incomplete or reduced penetrance** of disease mutations where for example in one family sharing a common germline mutation, several members display varying symptoms and ages of onset while others are protected from the disease entirely (Hahn et al., 2011, Bodor et al., 2012). What is the likelihood of developing overt disease or indeed providing a protection mechanism from disease in those "silent" germline mutation carriers? Thus far, we have only scratched the surface of the patterns and order of secondary genetic abnormalities that may contribute to such clinical variability and therefore the molecular basis of this phenomenon merits further investigation.

Going forward, not only is understanding the biology and genetic complexity of these rare inherited syndromes important for the families themselves but it can also offer valuable insights into the molecular pathogenesis of sporadic disease. It is hoped that by raising clinical awareness

of the known disease symptoms and maximising research endeavors on a collaborative basis we can improve the way we identify, test and manage these patients and their families and realise their clinical needs towards better outcomes. And indeed, a familial MDS/AML programme initiative, led by my supervisor Jude Fitzgibbon in collaboration with Inderjeet Dokal and Tom Vulliamy, started in 2015 with an untapped treasure trove of unique familial cases and an ambition to tackle these key research questions and this is where my PhD project is embedded.

In this thesis, recent advances in our understanding of the genetic basis of familial leukaemia and myelodysplasia will be discussed, including some of the challenges the field is facing in relation to defining novel germline mutations and optimal diagnostic practices for patients, before focusing on the molecular mechanisms underlying reduced penetrance of germline *GATA2* mutations, one of the key transcriptional regulators frequently mutated in inherited leukaemias. However, to set the scene, an overview of MDS/AML: its evolving molecular landscape and clinic-pathological and prognostic features, is warranted.

1.1 Sporadic MDS/AML

1.1.1 Incidence and Clinical Features

AML represents one of the commonest forms of acute leukaemias in adults, accounting for ~32% of all leukaemias and affecting approximately 3100 new cases each year in the United Kingdom (UK) (Cancer Research UK (CRUK)). It is a clinically diverse and genetically heterogeneous stem cell clonal disorder characterised by an uncontrolled proliferation of malignant immature myeloid progenitor cells (myeloblasts) within the bone marrow (BM) and the inability of these cells to fully differentiate into mature myeloid cells (**Figure 1.1**) (Grimwade et al., 2016). It can arise as a separate disease entity or preceded by a precursor myeloid malignancy, such as myeloproliferative neoplasms (MPN) or MDS. The latter is defined by peripheral blood (PB) cytopenias and morphologic and cytogenetic abnormalities of the BM (Invernizzi et al., 2015). According to CRUK (https://www.cancerresearchuk.org/about-cancer/acute-myeloid-leukaemia-aml), the incidence of MDS and AML increases with age, with median age of onset approximately 65 years old. Patients predominantly present with symptoms suggestive of BM failure and low blood counts; namely fatigue or breathlessness (due to anaemia), recurrent or severe infections (due to neutropenia) or a bleeding propensity (due to thrombocytopenia) (Grimwade, 2012).

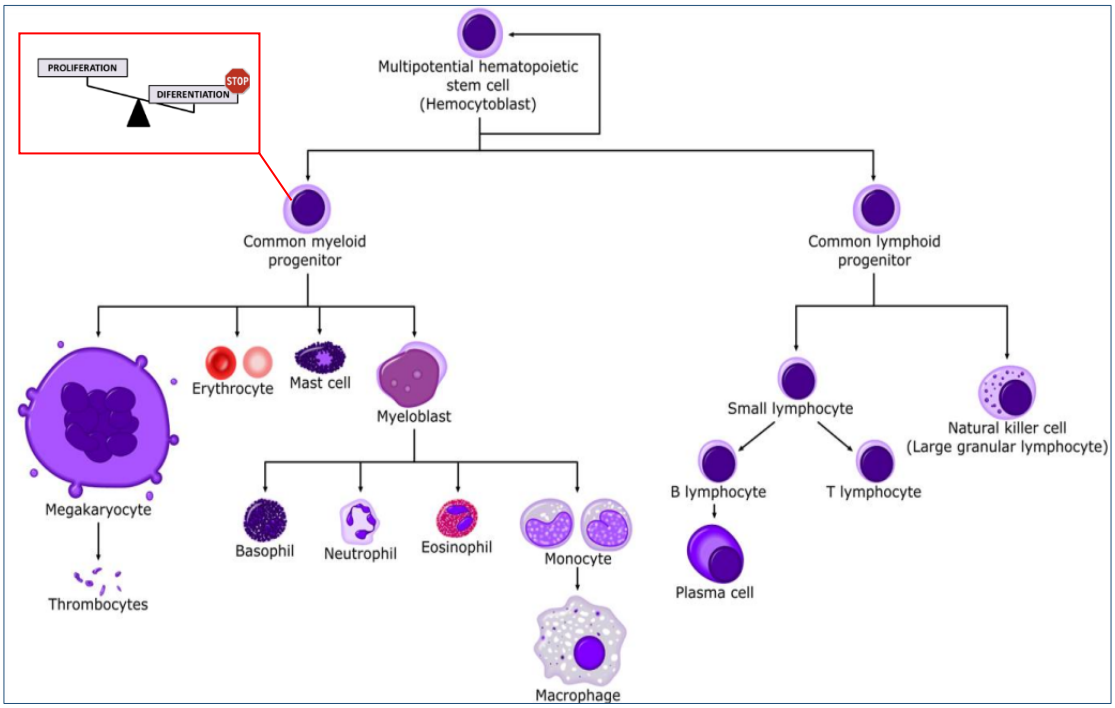


Figure 1.1 Schema of normal haematopoiesis in the bone marrow, where all blood cells have an equilibrium between proliferation and differentiation. AML typically occurs when this equilibrium is disrupted (as depicted in the red box) as malignant immature common myeloid progenitor cells start to proliferate abnormally in the bone marrow and differentiation is blocked as a result. Figure reproduced from (Shaikh and Bhartiya, 2012).

A suspicion of AML is primarily made on the basis of an abnormal full blood count followed by examination of a blood film to assess for the presence and percentage of leukaemic blasts. The diagnosis is typically confirmed on a BM aspirate and/or a trephine biopsy looking at cell numbers, morphology and BM architecture. Furthermore, immunophenotyping (immunohistochemical and flow cytometric), cytogenetic and molecular genetic analyses are conducted in parallel, after which tests results are integrated and optimal treatment options are considered on a case-by-case basis as part of a multidisciplinary meeting. Samples are also taken to identify a leukaemia specific phenotype or a molecular marker for the assessment of minimal residual disease (MRD), which involves detecting the persistence of AML post-remission and represents an important prognostic marker for relapse (Grimwade and Freeman, 2014).

1.1.2 Molecular Classification, Prognostic Factors and Cytogenetic Risk Groups

"*Cancer was not disorganized chromosomal chaos. It was organized chromosomal chaos...*"

— Siddhartha Mukherjee, The Emperor of All Maladies

Leukaemias are genetically simple and the relatively easy access to malignant cells has made it the forerunner of molecular-based discoveries. Forty years ago, the pioneering work of Janet Rowley and others uncovered the importance of somatic chromosomal abnormalities including balanced translocations (e.g. t(8;21), t(9;22) and t(15;17)) giving rise to in-frame chimeric fusion and target genes encoding haematopoietic transcription factors (e.g. *RUNX1*, *RARA*), epigenetic regulators (e.g. *MLL* (*KMT2A*)) and components of the nuclear pore complex (e.g. *NUP98*) (Rowley, 1980, Dohner et al., 2017, Rowley, 1973).

These molecular discoveries paved the way towards the introduction of disease classification systems, the most widely known schema is the World Health Organization (WHO) classification, providing an important framework for diagnosis, patient risk stratification and therapy decision making. In 2001, WHO published the 'Classification of Tumours of the Hematopoietic and Lymphoid Tissues' which superseded its predecessors (e.g. the French-American-British (FAB) classification) by redefining the AML BM blast threshold to 20% (Jaffe et al., 2001). Further investigations of the molecular aberrations in AML prompted the revised WHO classification in 2008 and its current incarnation in 2016 (**Table 1.1**) (Vardiman et al., 2009, Swerdlow, 2008, Arber et al., 2016). This integrated framework encompasses morphological, chromosomal and molecular findings and remains part and parcel of all multidisciplinary AML meetings today. Notably, the inclusion of familial myeloid malignancies as a separate disease entity in the 2016 WHO classification represents an important step in the recognition and management of this group of patients (discussed in section **1.2**).

Table 1.1 The 2016 WHO Classification of AML. Adapted from (Arber et al., 2016).

1) AML and Related Neoplasms	
AML with Recurrent Genetic Abnormalities	- AML with t(8;21)(q22;q22); *RUNX1-RUNX1T1* - AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* - APL with t(15;17)(q22;q12); *PML-RARA* - AML with t(9;11)(p22;q23); *MLLT3-MLL* - AML with t(6;9)(p23;924); *DEK-NUP214* - AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); *RPN-EVI1* - AML (megakaryoblastic) with t(1;22)(p13;q13); *RBM15-MKL1* - Provisional entity: AML with *BCR-ABL1* - AML with mutated *NPM1* - AML with mutated *CEBPA* - Provisional entity: AML with mutated *RUNX1*
AML with MDS-related Changes	
Therapy-related Myeloid Neoplasms	
AML, Not Otherwise Specified	- AML with minimal differentiation - AML without maturation - AML with maturation - Acute myelomonocytic leukaemia - Acute monoblastic/monocytic leukaemia - Acute erythroid leukaemia i) Pure erythroid leukaemia ii) Erythroleukaemia - Acute megakaryoblastic leukaemia - Acute basophilic leukaemia - Acute panmyelosis with myelofibrosis
Myeloid Sarcoma	
Myeloid Proliferations Related to Down Syndrome	- Transient abnormal myelopoiesis - Myeloid Leukaemia associated with Down Syndrome
Blastic/Plasmacytoid Dendritic Cell Neoplasm	
2) Myeloid Neoplasms with Germline Predisposition	
3) Acute Leukaemias of Ambiguous Lineage	
- Acute undifferentiated leukaemia - Mixed phenotype acute leukemia with t(9;22)(q34.1;q11.2); *BCR-ABL1* - Mixed -phenotype acute leukemia with t(v;11q23.3); *KMT2A* rearranged - Mixed phenotype - acute leukemia, B/myeloid, NOS - Mixed phenotype acute leukemia, T/myeloid, NOS - Provisional entity: Natural killer (NK) cell lymphoblastic leukemia/lymphoma	

1.1.2.1 Cytogenetic Risk Classification

Recurrent chromosomal rearrangements remain the most established prognostic markers in AML, dividing patients into 3 outcome groups according to favourable, intermediate and poor risk cytogenetic abnormalities (**Table 1.3**) (Grimwade et al., 1998, Grimwade et al., 2010, Grimwade, 2012). The favourable risk group includes t(8;21) translocations, inv(16)/t(16;16) and the t(15;17) *PML-RARA* fusion gene, whilst the poor risk group represents ~17% of patients with AML and includes those with monosomies or deletions of chromosomes 5 and/or 7, chromosome 3q or 11q abnormalities or with complex karyotypes (≥4 unrelated abnormalities). The 10-year survival in this patient group is dismal (<10%) necessitating improvement on existing treatments and care (Dombret and Gardin, 2016). Indeed, this is an active research area in my host laboratory who are performing multi-omic profiling to capture the multi-faceted personality of this poor risk disease, with the hypothesis that improvement in patient care will not be achieved by focusing on a single deregulated pathway or an actionable mutation. Lastly, the intermediate risk category represents ~50% of AML cases, including all normal karyotype AML (NK-AML) and those not otherwise classified as favourable or poor risk. Clinical outcomes in intermediate risk patients are heterogeneous with long-term survival estimated at 35-40% in young adults (Rockova et al., 2011, Breems et al., 2005). This has prompted the quest for additional prognostic markers to further stratify the disease and offer a more accurate prognosis for patients in this group.

Collectively, whilst cytogenetic risk group rarely influences the choice of induction therapy (as treatment usually begins before these results are available), it has proven powerful for informing optimal consolidation strategies, in particular, whether a patient should undergo allogeneic HSCT or not (Bullinger et al., 2017). The current degree of uncertainty around patients with intermediate cytogenetic risk means that a proportion of patients who would have been cured

from chemotherapy alone will go on to have a transplant with its associated morbidity, and improved prognostic scoring is therefore warranted in order to target this potentially toxic therapy for only those patients who really need it.

AML that arises from a pre-leukaemic MDS has the worst prognosis. Indeed, chromosomal alterations are observed in more than 50% of MDS cases, including deletions of chromosomes 5q, 7q and 20q, trisomy 8 and monosomy 7 (Cazzola et al., 2013a). The International Prognostic Scoring System (IPSS-R) is the most used tool for MDS classification (**Table 1.2**). The overall risk score is based on variables including patient's age, BM blast percentage, karyotype and white-cell counts, all of which can be used to predict overall survival (OS) and likelihood of AML transformation (Greenberg et al., 2012).

Table 1.2 IPSS-R classification criteria of MDS. Adapted from (Greenberg et al., 2012).

Risk Category	Cytogenetic Abnormality
Very Good	-Y, del(11q)
Good	Normal, del(5q), del(12p), del(20q), double including del(5q)
Intermediate	del(7q), +8, +19, i(17q), any other single or double independent clones
Poor	-7, inv(3)/t(3q)/del(3q), double clone including -7/del(7q), Complex: 3 abnormalities
Very Poor	Complex: >3 abnormalities

1.1.2.2 Identification of Molecular Markers with a Prognostic Impact

Over the last twenty years, research efforts have focused on deciphering the molecular basis of AML (particularly NK-AML), leading to the discovery of recurrent mutations in three genes: *FLT3*, *NPM1* and *CEBPA*, all of which are considered the earliest prognostic markers in AML that define survival outcomes (Schlenk et al., 2008, Vardiman et al., 2009) (**Table 1.3**). For example, while *FLT3* mutations modify risk from intermediate to poor and lead to reduced OS, *NPM1* and biallelic *CEBPA* mutations have been linked with favourable prognosis and long term OS approaching 60% (Dufour et al., 2010). These genes are now recognised as an essential component in the AML risk stratification system, leading to their incorporation into the revised WHO classification as confirmed disease entities and subsequently, into routine clinical testing (Arber et al., 2016, Dohner et al., 2017, Vardiman et al., 2009) .

Table 1.3 The 2017 European Leukemia Net (ELN) prognostication of AML risk groups. Adapted from (Dohner et al., 2017).

Risk Category	Cytogenetic/Molecular Genetic Abnormality
Favourable	t(8,21) (q22;q22); *RUNX1-RUNX1T1* inv(16) (p13.1;q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* t(15;17)(q24;q21) *PML-RARA* Mutated *NPM1* without *FLT3-ITD* (normal karyotype) Biallelic mutated *CEBPA* (normal karyotype)
Intermediate	Mutated *NPM1* and *FLT3-ITD* (normal karyotype) Wildtype *NPM1* and *FLT3-ITD* (normal karyotype) Wildtype NPM1 without FLT3-ITD (normal karyotype) Normal karyotype not classified as favourable. t(9;11)(p22;q23); *MLLT3-KMT2A* Cytogenetic abnormalities not classified as favourable or poor
Poor	inv(3)(q21q26.2) or t(3;3)(q21;q26.2);*GATA2-MECOM* (*EVI1*) t(6;9)(p23;q34);*DEK-NUP214* t(v;11)(v;q23); KMT2A rearranged t(9;22)(q34.1;q11.2); *BCR-ABL1* -5 or del(5q); -7; -17/abn(17p) Complex karyotype (>3), monosomal karyotype Wild type NPM1 and FLT3-ITD Mutated *RUNX1* Mutated *ASXL1* Mutated *TP53*

1.1.3 The Genetic Landscape of MDS and AML

> *"The capacity to blunder is the real marvel of DNA.*
> *Without this special attribute, we would still be anaerobic bacteria and there would be no music"* 
>
> — Lewis Thomas, The lives of a cell. Notes of a biology watcher

Aside from the earliest three genes discovered, recent advancements in NGS technologies have pushed the boundaries of unbiased AML genomic discoveries even further by capturing the enormous genetic heterogeneity. This has led to the characterisation of a plethora of genes and pathways known to have a role in haematopoiesis with the potential to inform classification and selection of therapies. Notably, AML and MDS were shown to share molecular lesions occurring due to deregulation in pathways involving: RNA splicing machinery, transcriptional regulation, epigenetic modification, DNA repair, signal transduction and the cohesin complex (Bejar et al., 2011, Marcucci et al., 2011, Grossmann et al., 2012, Cazzola et al., 2013b, Grimwade et al., 2016, Papaemmanuil et al., 2016). Indeed, there are now >100 genes known to be recurrently mutated in AML and MDS albeit many occur at low frequencies (<5%). **Table 1.4** shows a list of the top 25 mutated genes ranging in frequency from ~1-30%, and outlines the prognostic significance of these molecular lesions.

Perhaps one of the key findings from these NGS studies was the identification of recurrent actionable mutations affecting components of the epigenetic machinery, resulting in the co-ordinated disruption of transcriptional control and chromatin remodeling. Such mutations exert their effects via alterations in histone structure e.g. methylation, ubiquitination and phosphorylation which in turn lead to global alterations in gene expression. Prior to the introduction of NGS, *MLL* partial tandem duplications (PTDs) were the sole recurrent epigenetic lesions and cumulatively, additional epigenetic lesions have been identified. These are enriched

in genes encoding regulators of DNA methylation (*DNMT3A*, *TET2, IDH1, IDH2*) and post-translational histone modifications (*ASXL1, EZH2*), majority of which are loss-of-function aberrations (Delhommeau et al., 2009, Ley et al., 2010, Ntziachristos et al., 2016)(**Table 1.4**). These findings challenge the theoretical "two-step" model of leukaemogenesis (Gilliland, 2002), whereby the pathogenesis of leukaemia is not merely driven by mutations that confer a proliferative and/or survival advantage (e.g. *FLT3* and *c-Kit*) and a block in differentiation (*NPM1*, *RUNX1* and *CEBPA*). Instead, they form part of a bigger picture, further demonstrating the beauty and complexity of the (epi)genetic make-up of these diseases.

Table 1.4 Key pathways and genetic lesions recurrently mutated in AML. Data derived from (Papaemmanuil et al., 2016) and (Cancer Genome Atlas Research, 2013).

Functional Category		Gene	Mutation Frequency in AML (%)	Prognostic Impact
Nuclear Regulators		*NPM1*	28.3	Favourable in NPM1-mutant without FLT3-ITD
Cell Signalling		*FLT3*	31	Adverse
		NRAS	12.5	*conflicting information*
		PTPN11	5.9	Adverse in *NPM1:DNMT3A* mutant
		KRAS	4.6-10	*conflicting information*
		c-KIT	4.4	Adverse in CBF-AML
Transcription Factors		*RUNX1*	11.3	Adverse
		TP53	7.8	Adverse
		CEBPA	7.4	Double-mutant: favourable
		WT1	5.7	*conflicting information*
		GATA2	4	*conflicting information*
Epigenetic modifiers	**DNA Methylation**	*DNMT3A*	23.4	Adverse
		TET2	9.4	Adverse
		IDH2	9.7	*conflicting information*
		IDH1	8.6	*conflicting information*
	Histone Modifications	*MLL*	5.3	Adverse
		ASXL1	4	Adverse
		EZH2	2.3	Adverse in MDS
		BCOR	1.8	Adverse in MDS
Cohesin Complex		*SMC3*	4	*conflicting information*
		STAG2	3.8	*conflicting information*
		RAD21	2.7	*conflicting information*
Spliceosome Complex		*SRFS2*	3.2	Adverse in MDS
		U2AF1	2.8	Adverse in MDS
		SF3B1	1.6	Favourable in MDS

1.1.3.1 Improved Genomic Classification

Over the past few years, applications of NGS have led to further refinement of AML disease classification and prognostic stratification. Taking together findings from different sequencing studies, two new provisional entities: AML with mutated *RUNX1* and AML with *BCR-ABL1* have now been included in the current WHO classification (**Table 1.1**) and mutations in three genes (*ASXL1, RUNX1* and *TP53*) have been incorporated into the risk stratification algorithm as new molecular markers that confer a poor prognosis (Arber et al., 2016) (**Table 1.3**). However, this is just the beginning; a recent study by Papaemmanuil et al. (2016) defined the driver genomic landscape further in more than 1,500 patients with AML. This comprehensive analysis revealed four new AML molecular subgroups including *NPM1* emerging as a separate entity, rarely found in isolation and for which its prognostic impact is largely dependent on concurrent mutation(s). Remaining subgroups include AML with mutated chromatin or spliceosome genes, AML with *TP53* mutations, chromosomal aneuploidy or both and AML with $IDH2^{R172}$ mutations (Papaemmanuil et al., 2016). By implementing these classification schemes, patients with AML can be segregated into several distinct subgroups based on their underlying genetic abnormalities, allowing for more stratified approach to treatment and management.

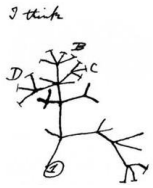
1.1.3.2 Patterns of Mutational Co-occurrence Modifying Clinical Outcomes

It is worth noting that not all of these mutations occur independently and non-random molecular associations often take place and have the potential to shape prognosis (Papaemmanuil et al., 2016, Cancer Genome Atlas Research, 2013). For example, extensive analyses from the cancer genome atlas (TCGA) consortium sequencing ~200 sporadic AML cases reported an average of 13 coding mutations per genome, of which recurrent mutations in *DNMT3A*, *IDH1/2*, *RUNX1* and *CEBPA* were found to be mutually exclusive with transcription factor fusion genes, suggesting the former were putative initiating lesions (Cancer Genome Atlas Research, 2013).

In contrast, molecular co-operation is frequently noted between *FLT3-ITD* and *NPM1* mutations (Gale et al., 2008, Grimwade et al., 2016), CBF translocations and *c-KIT* mutations (Paschka et al., 2006) and between *TP53* mutations and chromosomal aneuploidies, predicting a very poor outcome (Metzeler et al., 2016). Several studies also reported somatic *GATA2* mutations to co-occur with biallelic *CEBPA* mutations in approximately 20-40% of *CEBPA*-mutated sporadic AML cases and were associated with a favourable prognosis (Greif et al., 2012, Fasan et al., 2013b, Green et al., 2013). These studies shed light on the notion of predestination (canalisation), where the sequential order of mutation acquisition is essential in governing tumour initiation and evolution (more examples of familial cases is provided in section **1.2.2**).

1.1.4 Clonal Evolution in AML

> *"Whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved".*
>
> – Charles Darwin, On the Origin of Species



We now know that AML cell populations also undergo Darwinian evolution during their malignant transformation (Ferrando and Lopez-Otin, 2017). Indeed, an important revelation emerging in AML and cancer as a whole is that seldom does a biopsy represent a homogeneous population of tumour cells but rather, it is composed of a mosaic of multiple genomes, reflecting the expansion of subclones that confer a selective advantage and compete with progenitor clones for survival (Graubert and Mardis, 2011, Zhang et al., 2016). This concept was first introduced by Nowell in 1976 who postulated that cancer cells evolve and accumulate genetic and epigenetic alterations over a protracted period of time, some of which provide a survival advantage and subsequently lead to this "Darwinian" natural selection that promotes clonal

expansion, tumour progression and resistance to cytotoxic therapies (Nowell, 1976, Yates and Campbell, 2012).

Forty years later, and following the publication of the first AML genome by Ley et al. (2008), subsequent NGS studies have led to the development of novel computational algorithms with which to study AML clonal architecture and evolution by analysing the mutant allele fraction (or variant allele frequency (VAF)) of clonal cell populations (Welch et al., 2012). These studies revealed that mutational patterns seem to follow specific temporally ordered trajectories from which evolutionary phylogenetic trees can be inferred. For example, mutations in genes involved in epigenetic regulation such as *ASXL1*, *DNMT3A, IDH1/2* and *TET2* represent initiating clonal events (i.e. the trunk of the tree) with seemingly high VAFs. In contrast, mutations involving *FLT3* or *NPM1* represent secondary genetic events that occur later in leukaemogenesis (i.e. the branches of the tree) and therefore are likely to have lower VAF values (Bullinger et al., 2017).

These findings support the concept that stepwise acquisition of genetic events is required to fashion disease development. For example, initiating mutations (e.g. *DNMT3A, TET2 or IDH1/2*) can follow different routes of transformation depending on the secondary mutational event: *NPM1* mutations would lead to AML, *JAK2* mutations to myeloproliferative neoplasms and *SF3B1* mutations to MDS with ring sideroblasts (Papaemmanuil et al., 2016). Another example is an elegant study by Ding and colleagues who employed NGS to trace the clonal evolution of AML by comparing the mutational reservoir between paired diagnostic and relapse patient samples. They proposed two scenarios of disease relapse: a linear pattern, whereby relapse is driven by the evolution of the founding clone, or a branching evolution, where fitter subclones acquire additional mutations and expand leading to disease recurrence (Ding et al., 2012, Tawana et al., 2013).

It has also become apparent that some of these somatic clones might not always lead to clonal expansion and overt disease development. Instead, they can remain dormant in a pre-leukaemic or a quiescent state, and the term Clonal Haematopoiesis of Indeterminate Potential (CHIP) has been proposed to describe the phenomenon where certain recurrent mutations (e.g. *ASXL1*, *DNMT3A*, *TET2* and *TP53*) can be detected in the peripheral blood of elderly (>65 years), seemingly healthy individuals without prior diagnosis of a haematological malignancy (Becker, 2016). This suggests that there is a period of latency that precedes AML during which monitoring individuals should be considered to detect signs of progression to overt malignancy when these dormant clones have reawakened and start to proliferate aberrantly (Desai et al., 2018).

Nevertheless, a key challenge from these studies remains in discriminating driver from innocuous passenger or bystander mutations and further contextualise the prognostic impact of these molecular aberrations on disease pathogenesis. This is still a field very much in its infancy and work is now focusing on characterising tumour architecture and complexity in more detail, by using xenograft models of purified haematopoietic stem cells (HSC) and in-depth genetic profiling technologies such as single-cell sequencing with a view of dissecting the precise molecular lesions governing sensitivity vs. resistance to chemotherapy. Such findings will ultimately guide the future design and execution of targeted therapies to reduce the risk of relapse, the leading cause of death in AML (Tawana, 2015).

1.1.5 Current and New Treatment Strategies in MDS and AML

> *"Cancer is an expansionist disease; it invades through tissues, sets up colonies in hostile landscapes, seeking "sanctuary" in one organ and then immigrating to another. It lives desperately, inventively, fiercely, territorially, cannily and defensively — at times, as if teaching us how to survive…"*
>
> 
>
> — Siddhartha Mukherjee, The Emperor of All Maladies

The need to combat these leukaemias has never been more pressing. However, despite progressive improvement in our understanding of the molecular heterogeneity underlying these diseases, therapeutic strategies in AML have not significantly changed in over 40 years. Generally, the standard approach to treatment consists of two cycles of intensive induction chemotherapy to induce complete remission (CR), most commonly 7 days of cytarabine and 3 days of daunorubicin (so-called 7+3 regimen) or a variation, followed by a consolidation regimen depending on the patient's prognostic factors (e.g. age, co-morbidities, disease status following induction and genetic risk) (Yates et al., 1973, Dohner and Paschka, 2014). For example, patients with favourable prognosis undergo additional cycles of chemotherapy whilst those at a high-risk of relapse or with complex karyotype may benefit from an allogeneic HSCT given the availability of a suitable stem cell donor. Current treatments for low-risk MDS entail growth factor support, mainly erythroid stimulating agents (ESA) such as granulocyte-colony stimulating factor (G-CSF) and disease modulating agents like lenalidomide (List et al., 2005). In high-risk MDS, however, DNA hypomethylating agents such as azacitidine or decitabine have demonstrated clinical efficacy in patients who cannot tolerate allogeneic HSCT (Bejar and Steensma, 2014).

In the UK, the majority of patients are enrolled in national clinical trials, which may include randomisation to receive investigational therapies considered to be of potential benefit from earlier phase I and II trials. Furthermore, advances in donor selection techniques (human leukocyte antigen (HLA) typing) combined with improvements in the management of graft vs.

host disease (GvHD) and widespread adoption of reduced intensity conditioning protocols have led to reductions in transplant related mortality (TRM) (Horan et al., 2011). However, despite progress in therapeutics and supportive care, AML continues to pose a major clinical burden as many cases (especially elderly patients >60 years with 5-year survival estimated at 14%) experience short remission periods and ultimately relapse and develop resistance to standard chemotherapy, fully warranting the development of new therapeutic strategies (Buchner et al., 2009). Moreover, elderly patients with significant co-morbidities are often not able to withstand standard intensive chemotherapy regimens due to high treatment related mortality, leaving them with a negligible chance of cure and effective treatment options with reduced toxicity are therefore urgently needed. This can only be afforded by improving the diagnostic and prognostic tools at our disposal and increasing our understanding of the biology underlying this disease, uncovering new pathways and investigating ways in which these pathways can be therapeutically exploited.

Perhaps one of the most important implementations of NGS-based profiling directly relating to patient care is the identification of a specific pathway or mutation in AML that could define response (or a lack thereof) to targeted therapies. The most successful example has been the use of all-*trans* retinoic acid (ATRA) in *PML-RARA* positive cases, changing this disease from one with a dire prognosis to the highest cure rate of all AML subtypes and it is a key research aim to replicate the success of this agent in other types of AML (Castaigne et al., 1990, Kanamaru et al., 1995).

Indeed, giant strides are currently being made in unravelling the biological consequences of mutations discovered through large AML sequencing consortia, opening up new opportunities for therapeutic intervention. Recently approved targeted therapies include *FLT3* tyrosine kinase inhibitors (e.g. midostaurin) in poor risk AML cases with *FLT3-ITD* mutations and *IDH2* inhibitors

(e.g. Enasidenib or Idhifa) in *IDH2*-mutant AMLs (Stone et al., 2017, Garcia and Stone, 2017, Kim, 2017, Stein et al., 2017). Other promising examples include drugs that influence epigenetic regulation of gene expression e.g. histone deacetylase (HDAC) inhibitors or hypomethylating agents (e.g. 5-azacytidine) in *TET2*/*DNMT3A*/*ASXL1*-mutant AMLs and DOT1L inhibitors in *MLL*-rearranged leukaemias, demonstrating clinical utility in ongoing clinical trials (Dombret et al., 2015, Cruijsen et al., 2014, Daigle et al., 2013, Stein and Tallman, 2015).

Ultimately, it is hoped that the future clinical management of AML patients will move away from the classic "one size fits all" formula and towards a more stratified (patient and disease) targeted approach. This long-term mission might finally become a reality through initiatives such as the Beat AML® Master Trial, a collaborative effort led by the Leukemia & Lymphoma Society (LLS) where each AML patient enrolled on the trial can be screened for actionable mutations and then assigned to a targeted therapy based on his/her underlying genetic profile (http://www.lls.org/beat-aml) (Tyner et al., 2018). While these initiatives are by no means a final solution, they offer a substantive opportunity for precision medicine.

1.1.6 Minimal Residual Disease (MRD) Detection

Another emerging area of potential risk adapted therapy is the monitoring of persistent AML-associated mutations after therapy (also known as MRD). The conventional definition of remission post therapy involves the achievement of specific haematological parameters (demonstrating restoration of normal haematopoiesis) and the identification of <5% BM blasts (Ossenkoppele and Schuurhuis, 2016). Assays for the detection of MRD at levels not otherwise identified by standard morphology include real-time quantitative polymerase chain reaction (RT-qPCR) for the detection of fusion gene transcripts and molecular variants and secondly, multiparameter flow cytometry (MFC) for detecting leukaemia-associated immunophenotypes (LAIPs) (Ouyang et al., 2016).

Newer NGS advances such as single-cell sequencing offer the possibility of capturing disease heterogeneity at the single-cell level therefore allowing for more accurate assessment of MRD (Roloff et al., 2017, Ravandi et al., 2018).

The clinical application of MRD analysis at present is largely restricted to acute promyelocytic leukaemia (APL) and monitoring of t(15;17) *PML-RARA* transcripts (Tawana et al., 2013). In AML, however, MRD positivity of molecular aberrations (e.g. *NPM1*) has been linked with adverse outcomes (Grimwade et al., 2010, Schnittger et al., 2009, Ivey et al., 2016), indicating its potential efficacy as a prognostic marker for relapse. Opinions are often divided concerning the use and implementation of MRD analysis but it seems likely that it will prove particularly useful for select patients who will benefit from more aggressive therapy (e.g. allogeneic HSCT) in the first remission (Grimwade and Freeman, 2014) and in the absence of less effective therapies to target this residual population of leukaemia initiating cells (LICs) (Tawana, 2015, Ommem, 2016).

1.2 Familial MDS/AML

"The blood of your parents is not lost in you". 

– Menelaus, The Odyssey

1.2.1 The Current Molecular Landscape of Familial Leukaemia

While the majority of MDS/AML cases are indeed sporadic, less than 5% of cases are familial, where two or more individuals within the same family present with haematological malignancy (Godley, 2014). These rare cases are predominantly inherited in an autosomal dominant fashion, with median age of onset lower than sporadic disease and improving care and outcome in this high-risk group is essential, especially now that WHO listed hereditary myeloid malignancies as a separate disease entity in their 2016 revised classification system (Arber et al., 2016)

(**Table 1.1**). However, as alluded to already, the clinical recognition of inherited forms of these diseases is not an easy task; patients may be unaware of their predispositions, coupled with marked differences in disease latency and phenotype across mutation carriers and this is further complicated by a paucity of available family history and customised diagnostics guidelines and assays, leading to an under-representation of these familial cases (Rio-Machin et al., 2018b). Also, many mutation carriers remain asymptomatic into late adulthood due to incomplete or variable penetrance of mutations (discussed in section **1.3**) rendering their investigation and follow-up all the more challenging.

Such intra- and inter-familial disease heterogeneity highlights the high index of suspicion needed on behalf of the treating physician, to first recognise the symptoms and then perform appropriate testing to detect the underlying genetic predisposition. Notably, not only is studying familial leukaemias important for the individual families but it also offers insights into the management of sporadic disease. For example, lessons can be learned from familial leukaemia in relation to disease latency and penetrance, host genetics effects, the order by which mutations arise and novel disease mechanisms caused by genes only mutated in familial cases.

Following the identification of germline *RUNX1* mutations in 1999 by (Song et al.), advancements in NGS technologies have accelerated efforts to elucidate the genomic landscape of these inherited tumours. To date, leukaemia predisposition syndromes have been associated with germline mutations in ~14 discrete disease loci with variable mutation frequencies and clinical manifestations, illustrated in **Figure 1.2**. Our group at the Barts Cancer Institute have a longstanding clinical and research interest in this field and, together with other research groups worldwide, have made major contributions to the identification and discovery of families with germline mutations in these genes reported to date.

Unsurprisingly, some of these genes (e.g. *CEBPA* and *RUNX1*) are also recurrently mutated in sporadic leukaemia and are implicated in several key biological pathways and cellular processes including transcription, telomere maintenance and RNA processing, highlighting the considerable genetic heterogeneity. There is also clinical heterogeneity associated with each syndrome/genetic subgroup with a myriad of phenotypes varying from pre-existing platelet dysfunction to lymphoedema and BM failure (in addition to MDS and/or AML) (**Figure 1.2**) (Drazer et al., 2016, Akpan et al., 2018).

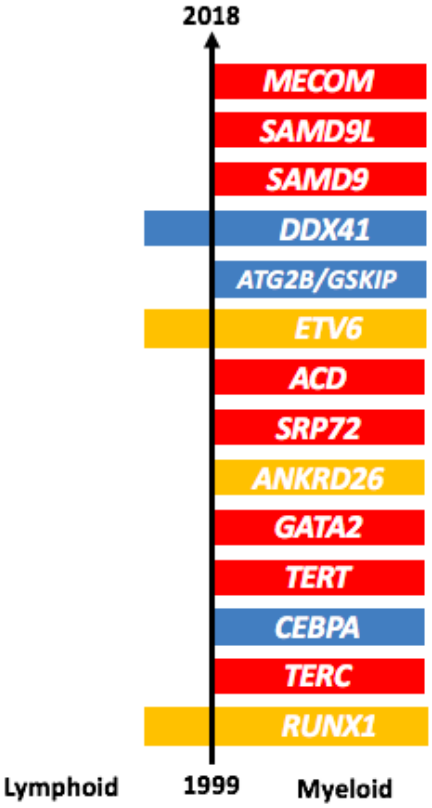


Figure 01.2 The genetic landscape of familial MDS/AML. The order of the 14 germline mutations is shown based on their date of discovery. Mutations are broadly assigned to 3 groups according to clinical phenotype: myeloid malignancies presenting 'purely' without a pre-existing disease (blue), accompanying abnormal platelet number/function (yellow) or associated with other organs dysfunction (red). These genes can also be classified according to their protein functions including transcription (*RUNX1*, *CEBPA*, *GATA2*, *ETV6*, *MECOM*), telomere maintenance (*ACD*, *TERT*, *TERC*), RNA processing (*DDX41*), cell trafficking (*SRP72*), inflammation (*SAMD9*, *SAMD9L*), and other unknown functions (*ANKRD26*, *ATG2B*). The most frequently mutated genes are *DDX41*, *GATA2* *RUNX1* and *TERT*. Germline *DDX41*, *ETV6* and *RUNX1* mutations predispose to both myeloid and lymphoid malignancies. Figure modified from (Tawana and Fitzgibbon, 2016).

The most well-recognised MDS/AML predisposition syndromes are: familial platelet disorder with propensity to develop AML (FPD/AML) due to mutations in *RUNX1*; pure familial AML with mutated *CEBPA*; *GATA2* deficiency syndromes and the inherited bone marrow failure syndromes (IBMFS) caused by mutations in genes including key telomere regulators (*TERC* and *TERT*) (West et al., 2014a).

These syndromes are briefly described in the ensuing section (**1.2.2**), concentrating on the unique features that distinguish them from one another, whilst *GATA2*, the focus of my thesis, is discussed in more detail in section (**1.2.3**). Lastly, an overview on how to identify, test and manage patients and families at risk for these inherited syndromes is provided in section (**1.2.4**), while also highlighting some questions arising in the field.

1.2.2 Familial MDS/AML Genetic Predisposition Syndromes

"Human beings are ultimately nothing but carriers – passageways – for genes. They ride us into the ground like racehorses from generation to generation".

– Haruki Murakami

1.2.2.1 FPD/AML with *RUNX1* Mutations

Familial platelet disorders are rare autosomal dominant diseases characterised by thrombocytopenia, clinical bleeding due to platelet dysfunction and a propensity to develop MDS and/or AML (Owen et al., 2008). Since the initial discovery of *RUNX1* germline mutations in 6 FPD/AML families by Song and colleagues in (1999), more than 30 families have been reported in the literature so far. These cases harbour a spectrum of *RUNX1* germline aberrations encompassing frameshift, missense or nonsense mutations localised to either of the two highly-conserved domains as shown in **Figure 1.3** (Liew and Owen, 2011, Nickels et al., 2013).

RUNX1 (also known as *CBFA2* or *AML1*) encodes an alpha sub-unit of the core binding factor (CBF) complex and is a key transcription factor important in haematopoietic regulation and myeloid differentiation (Owen et al., 2008). Germline mutations in *RUNX1* typically cluster in the N-terminal RUNT homology domain (RHD), predicted to disrupt DNA binding and heterodimerisation with CBFβ subunit, rendering *RUNX1* haploinsufficient in these cases (Song et al., 1999, Preudhomme et al., 2009).

Germline lesions in the C-terminal transactivation domain (TAD), however, are less frequent and are likely to impair interactions with downstream regulatory targets like *PU.1.* Such mutations exert a dominant-negative effect thereby antagonising WT RUNX1 function (Michaud et al., 2002).

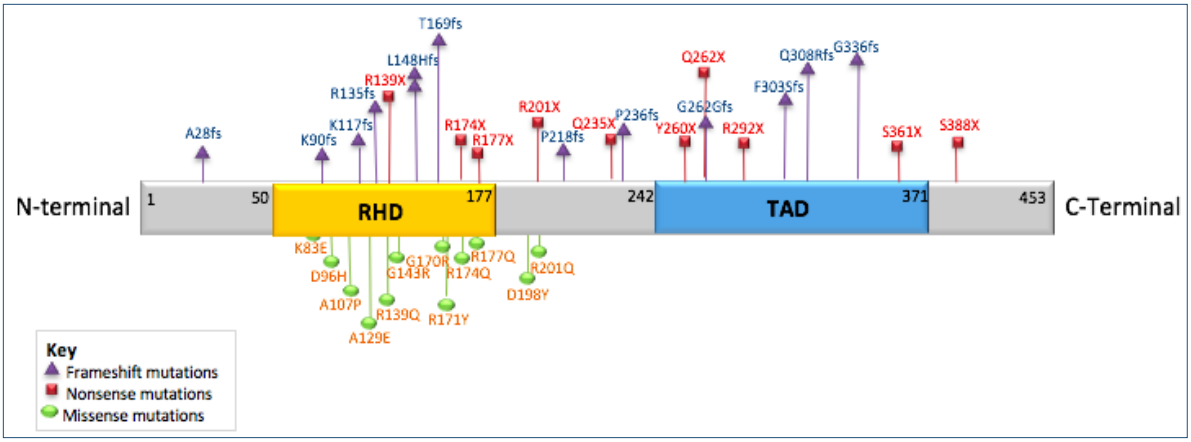


Figure 1.3 Schematic representation of the RUNX1 protein demonstrating the distribution of germline *RUNX1* mutations in familial FPD/AML reported to date. RHD: Runt homology domain; TAD: transactivation domain. Figure adapted from (Nickels et al., 2013).

There appears to be however no discernable differences in the clinical presentation between *RUNX1* dominant-negative mutations or those predicted to cause haploinsufficiency, although a higher frequency of MDS/AML might be associated with the former (Liew and Owen, 2011). However, several reports showed that ~20% of FPD/AML patients exhibit loss of one *RUNX1* allele by deletion of chromosome 21, leaving the remaining allele intact, suggesting haploinsufficiency of RUNX1 is also sufficient to cause FPD/AML (Beri-Dexheimer et al., 2008, Sakurai et al., 2016, Preudhomme et al., 2009). Germline testing analysis for individuals with suspected thrombocytopenia or FPD/AML should therefore not only be restricted to *RUNX1* mutations but also include tests sensitive to detect whole or partial chromosomal deletions or copy number alterations (CNAs) that may go unnoticed by more conventional sequencing techniques. These tests include comparative genomic hybridisation (CGH) arrays or multiplex

ligation-dependent probe amplification (MLPA) analyses (Jongmans et al., 2010). Indeed, a study by our group identified germline *RUNX1* deletions (ranging from 311-666Kb and mainly encompassing exons 1 and 2) across four FPD/AML pedigrees using array CGH and were further validated by MLPA ((Tawana, 2015) unpublished observations).

Moreover, a certain degree of heterogeneity in disease phenotypes exists within and between *RUNX1*-mutated families, rendering this syndrome difficult to recognise. Approximately 40% of *RUNX1* mutation carriers develop full-blown malignancy whilst others present with mild to moderate thrombocytopenia and some even retain normal PB counts (Owen et al., 2008). This suggests that germline mutations alone are not always sufficient to initiate disease and may need synergy with other molecular events to initiate clonal expansion and proliferation. This is exemplified by a study which reported acquisition of *CDC25C* variants as a recurrent or secondary molecular event in germline *RUNX1*-mutated Japanese families (Yoshimi et al., 2014). In addition, previous work from our lab performed comprehensive genetic profiling across multiple siblings from a Hungarian *RUNX1*-mutated pedigree with AML and showed somatic acquisition of variants upregulating JAK-STAT signalling pathway, including *JAK2* and *SH2B3* (a negative regulator of JAK2) whilst also sharing the 46/1 haplotype linked with sporadic *JAK2*-positive myeloproliferative neoplasms (MPNs). This study provided a notable example of intra-familial convergent AML evolution, where inherited genetic factors may govern somatic mutation acquisition and leukaemic transformation (Tawana et al., 2017b).

Other genetic lesions that confer thrombocytopenia and/or an inherited predisposition to haematological malignancy are *ANKRD26* (Noris et al., 2013) and *ETV6* (Zhang et al., 2015, Noetzli et al., 2015). The inter-relationship across different lesions is noteworthy, for example, *ETV6* is recurrently translocated with *RUNX1,* t(12;21), in ~25% of paediatric B-cell acute lymphoblastic leukaemia (ALL) patients (Romana et al., 1995, Ford et al., 1998).

1.2.2.2 Pure Familial AML with Mutated *CEBPA*

The transcription factor CCAAT/enhancer-binding protein alpha (*CEBPA*) represents another important loci predisposing to familial AML as its main presenting feature (without any precursor cytopenias or MDS). It is encoded on chromosome 19q13.1 and has an essential role in mediating granulocytic differentiation and cellular growth arrest (Nerlov, 2007).

While heterozygous mutations in this intronless, single-exon gene were reported in 10-15% of sporadic NK-AML, Smith and colleagues from the Fitzgibbon lab were the first to report germline *CEBPA* mutations in two siblings and their father who presented to St. Bartholomew's Hospital, London, two weeks apart with AML (Smith et al., 2004). All 3 family members harboured an identical N-terminal mutation (c.212delC:p.P23RfsX137). This frameshift lesion results in an increase in translation of the shorter p30 isoform, which lacks its first transactivation domain but nevertheless retains the leucine zipper region (bZIP) required for dimerisation with the p42 protein, leading to a dominant-negative inhibition (**Figure 1.4**). Disease progression in these siblings was accompanied by the acquisition of additional somatic in-frame insertions in the C-terminal section of the protein. Such mutations result in alteration of the structure of bZIP; dimerisation with WT CEBPA is therefore disrupted leading to a loss-of-function (LoF) (Smith et al., 2004, Pabst et al., 2008).

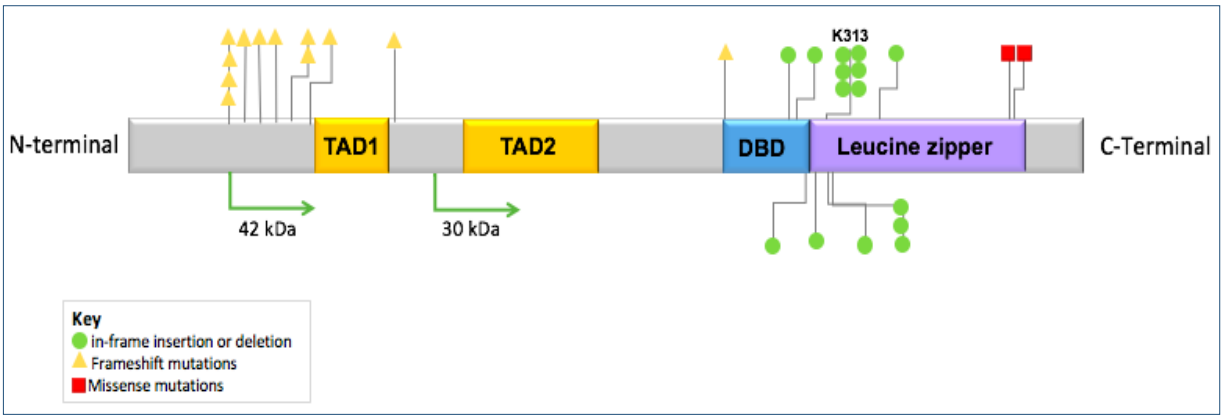


Figure 1.4 Schematic representation of the CEBPA protein demonstrating the distribution of germline and acquired *CEBPA* mutations in familial AML. The majority of germline mutations are located in the N-terminal region (frameshift) while somatic acquired mutations typically cluster within the C-terminal, with a hotspot located at residue K313. TAD 1/2: transactivation domain 1/2; DBD: DNA-binding domain. KDa: kilodalton. Figure adapted from (Nickels et al., 2013, Tawana et al., 2015).

Since this discovery, a retrospective analysis of 25 patients from 11 *CEBPA*-mutated pedigrees revealed that AML with germline *CEBPA* mutation is associated with favourable long-term clinical outcomes, with the molecular and clinical features resembling that of sporadic NK-AML with double-mutated *CEBPA* and a 10-year OS approaching 65% (Tawana et al., 2015, Dufour et al., 2010). However, despite this favorable prognosis, AML patients with germline *CEBPA* mutations are prone to the development of future leukaemias, occurring as late events (27 months). Indeed, a striking observation was made following deep sequencing of paired diagnostic and relapse samples from these patients revealing the occurrence of secondary somatic C-terminal *CEBPA* mutations at relapse that were absent at diagnosis. This suggests that not only were patients cured from their initial disease but rather the relapse was triggered by entirely new leukaemic episodes, supporting a new mode of disease recurrence and progression in familial AML. While these patients responded well to salvage therapies (i.e. are chemosensitive), allogeneic HSCT should be considered to alleviate the risk of AML recurrence, although the timing of HSCT (at the first or second remission period) has been controversial (Tawana et al., 2017a).

To investigate additional molecular events accelerating the development of AML, whole exome sequencing (WES) data showed clustering of identical acquired somatic mutations within 2 of these *CEBPA*-mutated families (C-terminal *CEBPA* with *GATA2* mutations and C-terminal *CEBPA* with *WT1* mutations and 11p acquired uniparental disomy (aUPD)) providing another evidence for convergent evolution whereby the nature of secondary mutations (or lack thereof) is prescribed by an individual's host genotype (Tawana et al., 2015).

While germline *CEBPA* mutations are predominately localised to the N-terminal domain and are associated with a near-complete penetrance albeit with a variable latency period (range 2-46 years), rare pedigrees with inherited C-terminal *CEBPA* mutations have recently been delineated and appear to demonstrate reduced penetrance, impeding clinical recognition and surveillance in those carriers (Pathak et al., 2016, Tawana, 2015) unpublished observations).

1.2.2.3 Inherited Bone Marrow Failure Syndromes (IBMFS) and Telomeropathies

IBMFS are a diverse group of disorders characterised by features indicative of BM failure with predisposition to hereditary MDS/AML. Although initially thought to be diseases of early childhood, these syndromes can also manifest in adulthood (West et al., 2014a) and range from Fanconi anaemia (FA) (caused by mutations in *FANCA* and *FANCD1*) to dyskeratosis congenital (DC) (mutations in *DKC1*), Shwachman-Diamond syndrome (SDS) and telomere biology disorders (TBD)(Bluteau et al., 2017). The latter is considered the most common IBMFS caused by heterozygous mutations in genes responsible for telomere maintenance: (*TERT,* encoding telomerase reverse transcriptase and *TERC,* encoding telomerase RNA component) so called the "telomeropathies" (Vulliamy et al., 2001, Yamaguchi et al., 2005, Townsley et al., 2014). The genetic and phenotypic complexity of telomeropathies has been further delineated by the discovery of mutations in other genes including *NOP10* (Walne et al., 2007), *TINF2* (Savage et al.,

2008) and *RTEL1* (Walne et al., 2013b, Cardoso et al., 2017). Such mutations lead to genomic instability, disruption of the DNA repair machinery and shorter telomeres (Kirwan et al., 2009, Dokal and Vulliamy, 2010).

More interestingly and specifically, mutations in *TERC* and *TERT* have been associated with anticipation, whereby younger generations present with more severe disease phenotypes and shorter telomeres compared to their older counterparts (Vulliamy et al., 2004). Awareness of such phenomenon and appropriate screening is essential as a child inheriting a *TERC* or *TERT* mutation could present even before his/her parent carrying the same mutation (West and Churpek, 2017). Ultimately, IBMFS are inherited in an autosomal dominant fashion (except for FA where the mode of inheritance is autosomal recessive) with a myriad of clinical manifestations and incomplete penetrance, highlighting the importance of integrating *TERT* and *TERC* into the diagnostic algorithm of familial MDS/AML and conducting telomere length testing in patients with suspected diagnosis of IBMFS.

- **Recently Identified Disease Genes in Familial MDS/AML (2015-2018):**

1.2.2.4 The RNA Helicase *DDX41*

Germline mutations in the RNA helicase DEAD/H-box polypeptide 41 (*DDX41*) have recently been associated with an inherited predisposition to MDS/AML and late age of disease onset (for certain mutations) similar to that of sporadic disease (mean 65.5 years). *DDX41* is located on chromosome 5q35.3 and defects in this gene lead to loss of tumor suppressor function due to altered pre-mRNA splicing and RNA processing, representing a new genetic class in familial MDS/AML. Polprasert et al (2015) initially reported germline *DDX41* mutations in seven MDS/AML families, five of which carried the recurrent frameshift (p.D140GfsX2) mutation and

subsequently, Lewinsohn and colleagues (2016) expanded this cohort even further by reporting nine additional MDS/AML families with novel germline heterozygous *DDX41* mutations. Notably, 50% of these affected members acquired somatic *DDX41* mutations (p.R525H) in the other allele, akin to biallelic *CEBPA*-mutated cases (Smith et al., 2004, Tawana et al., 2015, Tawana and Fitzgibbon, 2016). In the same year, four more MDS/AML families with novel germline *DDX41* variants and a tendency to shortened telomeres were described by our sister lab at the Blizard institute (Cardoso et al., 2016). Of note, somatic *DDX41* mutations are seldom seen without a germline *DDX41* alteration and therefore, identification of any *DDX41* mutation should raise a suspicion of a germline involvement (Brown et al., 2017). Notably, not only were *DDX41* mutations identified in families with MDS/AML but also with colorectal cancers and other haematological malignancies including non-Hodgkin's lymphoma (NHL), Hodgkin disease (HD) and multiple myeloma (MM) (Lewinsohn et al., 2016, Cheah et al., 2017).

The prolonged latency of MDS/AML associated with this genetic subtype may hinder the timely recognition of familial disease occurrence in older patients as well as appropriate donor selection when it comes to HSCT. This dilemma is demonstrated in two families (**Figure 1.5**) in which the index cases presented with NK-AML at age >40 years and both received HSCT from their respective, healthy HLA-matched brothers and subsequently relapsed with 99% donor chimericsm (Berger et al., 2017, Kobayashi et al., 2017). To find a genetic explanation for this disease recurrence, the authors from these two studies performed WES and showed the presence of germline heterozygous *DDX41* mutations in all brothers (donors and recipients). Targeted deep sequencing was then performed to detect co-operating somatic mutations that could contribute to AML development and identified acquired mutations in *DNMT3A*, *ASXL1* and even secondary somatic missense *DDX41* mutations (p.R525H) shared between donor and recipient samples. Altogether, these rare incidences indicate that disease relapse was likely of

donor origin (donor-derived leukaemia) and underline the necessity for heightened clinical awareness and comprehensive family screening (even in older members) to detect these mutations prior to making any donor selection related decisions.

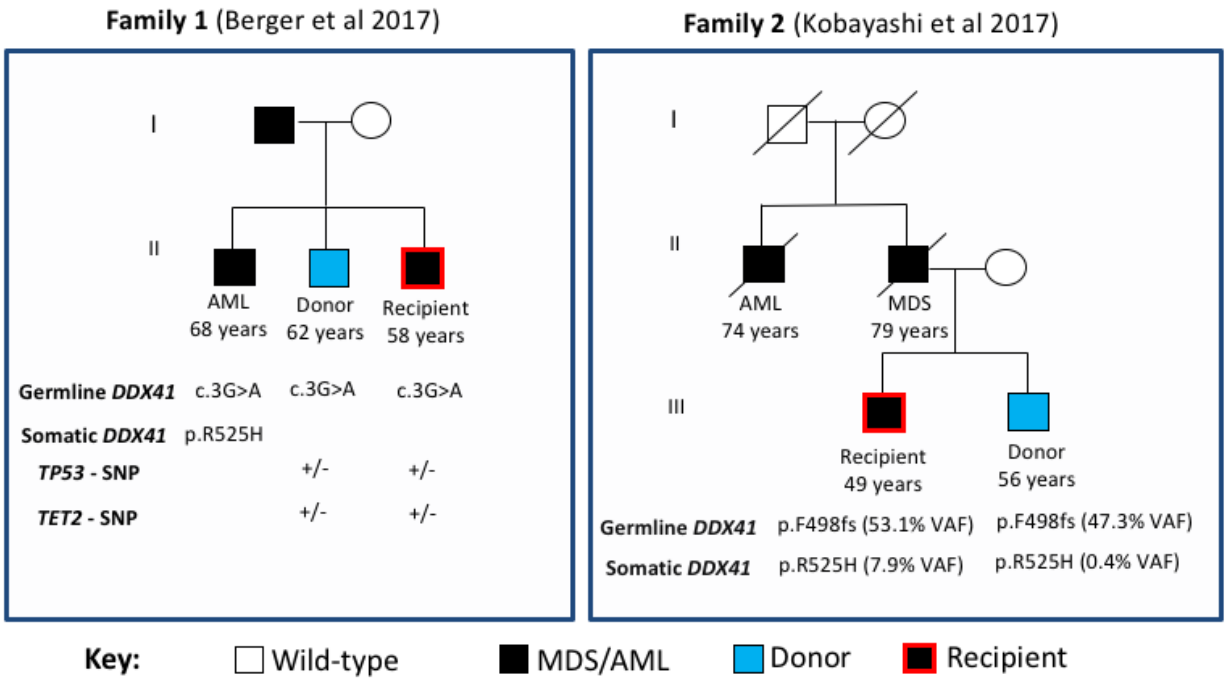


Figure 1.5 Donor-derived leukaemia in germline *DDX41*-mutated families.

1.2.2.5 *SAMD9, SAMD9L* and Making a Monosomy 7

Two newly discovered genes recently joined the catalogue of germline variants predisposing to MDS/AML. The *SAMD9L* locus and its paralog SAMD9, residing head to tail on chromosome arm 7q21, function to inhibit cell proliferation and are therefore regarded as tumor suppressor genes. Whilst *SAMD9* mutations were associated with a syndrome of MDS, infection, restriction of growth, adrenal hypoplasia, genital phenotypes and enteropathy (known as the MIRAGE syndrome), germline heterozygous gain-of-function (GoF) *SAMD9L* mutations can cause a syndrome of cytopenia, immunodeficiency, MDS and neurological symptoms

(e.g. Ataxia-Pancytopenia) (Narumi et al., 2016, Chen et al., 2016, Tesi et al., 2017, Pastor et al., 2018, Davidsson et al., 2018).

Consistent with a GoF effect, Tesi et al showed that *SAMD9L* missense mutants can cause decreased cell proliferation compared to WT protein, leading to loss of *SAMD9L*-mutant allele and cytopenias (Tesi et al., 2017). The resultant haematopoietic crisis can therefore facilitate selection and expansion of monosomy 7 clones, foreshadowing the development of overt MDS. On the other hand, some *SAMD9L* GoF mutation carriers can experience milder clinical presentation and recovery from their cytopenias explained by haematopoietic revertant mosaicism derived from either UPD of 7q or additional somatic *SAMD9L* LoF truncating mutations in *cis,* disguising carriers of *SAMD9L* germline mutations. Further studies are therefore needed to determine the likelihood of developing overt disease in those carriers. Overall, these studies highlight a novel mechanism in familial MDS/AML involving monosomy 7 driven by the selective loss of germline *SAMD9L* mutation. However, it seems likely that *SAMD9*/*SAMD9L* are yet another gene entity that behave differently in adults and children; germline LoF *SAMD9* and *SAMD9L* mutations have recently been identified in adult MDS and convey different pathophysiologic effects than their GoF counterparts (Nagata et al., 2018, Pastor et al., 2017).

1.2.3 Focusing on *GATA2* – the Stemness Gene

Transcription factors (TFs) are considered master regulators of haematopoiesis through their roles in orchestrating cellular proliferation and differentiation, and so it comes as no surprise that disruption of such control, through mutations or altered gene expression in these TFs, can trigger the development of overt malignancy in a Domino like fashion. Take GATA2, one of six GATA binding protein factors and key players in this ensemble, essential for the development and differentiation of haematopoietic and lymphatic vascular systems. It is expressed in early myeloid progenitors and plays a key role in cell fate determination and target gene regulation, joining the ranks of other TFs, namely *RUNX1* and *CEBPA* (Kazenwadel et al., 2012, Vicente et al., 2012, Collin et al., 2015).

1.2.3.1 Gene Structure, Role in Haematopoiesis and Transcriptional Regulation

The *GATA2* gene is mapped to chromosome 3q21.3, composed of six coding exons (the first exon is untranslated (UTR)) and encodes two highly conserved zinc-finger domains (ZF1 and ZF2) (**Figure 1.7**). Its name was derived owing to the ability of these domains to bind the consensus DNA sequence site (A/T)GATA(A/G) in promoter regions of target genes to enable accessible transcription. GATA2 is a 480 amino acids long (50.5kDa) and is predominantly expressed in early haematopoietic stem cells in the BM, hence it has been called the haematopoietic "stemness" gene (Scott et al., 2010). It is also expressed in mature megakaryocytes, monocytes and mast cells as shown in **Figure 1.6** (Hahn et al., 2011, Rodrigues et al., 2012, Vicente et al., 2012).

The function of GATA2 in normal haematopoiesis has been widely studied *in vivo*. For example, Tsai and colleagues showed that *Gata2* $^{-/-}$ mice succumb to anaemia at day 10 of gestation due to failure of definitive haematopoiesis, whereas mice heterozygously deficient for *Gata2*$^{-/+}$ exhibit dramatically reduced numbers of BM progenitor cells associated with increased

apoptosis and cellular quiescence, demonstrating its importance in the proliferation and survival of early stages of blood development (Tsai et al., 1994, Tsai and Orkin, 1997, Rodrigues et al., 2005). Precisely, GATA2 controls the transition from endothelium to haematopoietic stem cells (HSC). During HSC differentiation, GATA2 plays a critical role in controlling downstream cell fate decisions, working in concert with TFs CEBPA, GATA1 and SPI1 (PU.1). For example, GATA2 is displaced from chromatin by GATA1 (a process known as the GATA switch) which drives HSC towards erythroid and megakaryocytic differentiation (**Figure 1.6**) (Vicente et al., 2012, Collin et al., 2015, Wlodarski et al., 2017).

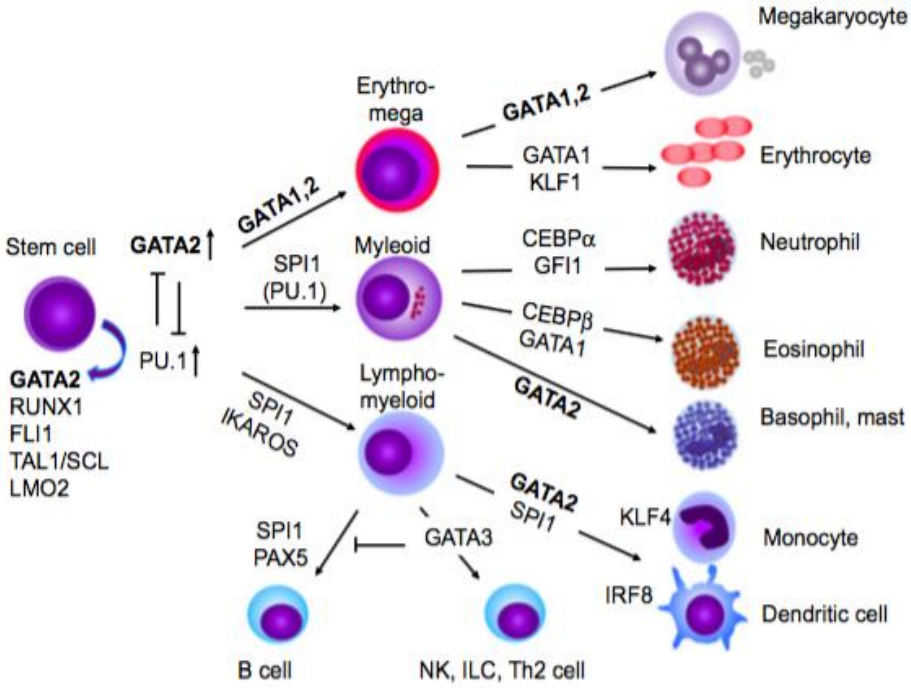


Figure 1.6 The role of GATA2 in haematopoietic development and differentiation.
A model of key interactions of GATA2 with other major lineage-specifying TFs.
Figure reproduced from (Collin et al., 2015).

In addition to its role in haematopoietic differentiation, GATA2 is involved in lineage-specific transcriptional regulation. Indeed, GATA2 collaborates with six other TFs (ERG, FLI1, LMO2, LYL1, RUNX1 and TAL1) forming a core heptad regulatory unit to direct early lineage development (**Figure 1.6**) (Vicente et al., 2012).

Interestingly, using chromatin immunoprecipitation (ChIP), it was demonstrated that not only is *GATA2* expression regulated by a cloud of proteins including BMP4, CEBPA, ETS1, EVI1, HOXA9 and NOTCH1 but also it regulates its own transcription by binding to its own promoter (Cortes-Lavaud et al., 2015, Vicente et al., 2012). Several reports showed that GATA2 modulates the expression of other target genes downstream such as *GATA1*, *LMO2*, *PU.1*, *RUNX1, SCL* and *TAL1* (Gao et al., 2015, Katsumura et al., 2017). Altogether, these findings highlight the importance of complex interactions and crosstalk between lineage-specific TFs in regulating haematopoiesis and how mutations in either of these genes can perturb the network and contribute to overt haematopoietic disease, including leukaemia.

1.2.3.2 Clinical Syndromes Associated with Germline *GATA2* Mutations – *GATA2* Deficiency

While *GATA2* is mutated in ~4-7% of sporadic MDS/AML, especially in 20-40% of biallelic *CEBPA*-mutated cases and typically cluster in the first zinc-finger (ZF1) domain (Greif et al., 2012, Fasan et al., 2013b, Green et al., 2013), the first report depicting the role of germline *GATA2* mutations in familial disease was in 2011. It started when Hahn and colleagues identified germline *GATA2* lesions in four families presenting with pure "out of the blue" leukaemia inherited in an autosomal dominant manner. All affected individuals harboured heterozygous *GATA2* mutations within the second zinc-finger (ZF2) domain. The missense mutation (p.Thr354Met) was identified in three families whilst a 3bp deletion (p.Thr355del) was described in the fourth family, all of which are regarded as LoF aberrations and predicted to impair DNA binding (Hahn et al., 2011). Shortly thereafter, more than 30 *GATA2*-mutated families were reported as outlined in **Table 1.5**.

Remarkably, not only were germline *GATA2* mutations associated with pure familial MDS/AML in those families but rather with a wide spectrum of overlapping clinical manifestations

encompassing immunodeficiency, vulnerability to mycobacterial, viral or fungal infections, warts, autoimmunity and cytopenias (Hyde and Liu, 2011, Hsu et al., 2013, Dickinson et al., 2011, Ostergaard et al., 2011). These fall under the umbrella of "*GATA2* deficiency" syndromes and include: Emberger syndrome (primary lymphoedema with predisposition to MDS/AML), MonoMAC syndrome (monocytopenia and mycobacterium infection) and DCML (dendritic cell, monocyte, B- and natural killer-lymphoid) deficiency, reinforcing the general notion that *"Every gene is pleiotropic, it influences multiple traits to varying degrees. Every trait is multigenic, multiple genes contribute to the expression of every phenotypic detail"* – Myers.

Table 1.5 Prevalence of inherited *GATA2*-mutated myeloid malignancies since its initial description in 2011. Adapted from (Wlodarski et al., 2017).

Year of Study	Reference (et al)	Clinical Phenotype	No. of *GATA2* -mutated cases	No. of cases with myeloid malignancy	Age at diagnosis; range (median) years
2011	Hahn	Familial MDS/AML	21	15	10-53 (20.5)
2011	Ostergaard	Emberger Syndrome	14	8	9-53 (12)
2012	Bödör	Familial MDS/AML	6	3	18-31 (24.5)
2012	Kazenwadel	Familial MDS/AML Lymphoedema	10	9	10-33 (16)
2012	Holme	Familial MDS/AML Lymphoedema	4	4	12-48 (20)
2013	Pasquet	Chronic Neutropenia	14	10	6-35 (15)
2013	Hsu	MonoMAC	32	20	3-78 (21.5)
2014	Spinner	*GATA2* Deficiency	57	42	0.4-78 (19)
2014	West	*GATA2* Deficiency	48	42	12-78 (35.5)
2014	Dickinson	DCML	30	11	4-40 (25)
2015	Mir	Familial MDS/AML	5	3	7-38 (22.5)
2015	Ganapathi	*GATA2* Deficiency	28	28	14-60 (30)
2015	Churpek	Familial MDS/AML	7	6	13-68 (16.5)

2015	Wang	Paediatric MDS	6	5	Not Available
2015	Zhang	Paediatric MDS	5	5	12-22 (16)
2016	Wlodarski	Paediatric MDS	60	57	3-19 (12)
2016	Novakova	Paediatric MDS	12	10	4.4-17 (14.5)
2017	Schlums	*GATA2* Deficiency	13	5	7-60 (18)
2018	Galera	Familial MDS/AML (Donor-derived AML)	6	5	13-64 (34)
Total			**378**	**288 (76%)**	**12-35.5 (19.7)**

It is worth noting here however that not all *GATA2*-deficient individuals exhibit all (or even any) of these features and their timing and sequence of manifestation are not fully understood (Horwitz, 2014). The spectrum of *GATA2*-related disease extends even further to include thrombosis, congenital neutropenia, deafness, pulmonary complications and bone marrow failure (Pasquet et al., 2013, Spinner et al., 2014). *GATA2* deficiency also appears to be key in the aetiology of paediatric MDS. Indeed, Wlodarski et al screened >600 cases of children and young adults with *de novo* MDS and showed that 7-15% of these cases carried germline *GATA2* mutations complicated by systemic infections and a high risk of developing MDS/AML (Wlodarski et al., 2016, Hirabayashi et al., 2017).

1.2.3.3 Landscape and Functional Consequences of Germline *GATA2* Mutations

The mutations identified in these aforementioned *GATA2* deficiency syndromes are scattered throughout the gene, which may explain the observed clinical heterogeneity, and include protein-truncating frameshifts, missense mutations and in-frame insertions or deletions (**Figure 1.7**) (Holme et al., 2012, Ostergaard et al., 2011). In particular, *GATA2* germline mutations in familial MDS/AML are predominantly missense and appear to cluster mainly within ZF2 domain, acting as a mutational hotspot which mediates interactions with other proteins and TFs such as

PU.1 and FOG-1 (Hahn et al., 2011). Several lines of evidence suggest that these *GATA2* mutations impair its transcription and promoter activation leading to a LoF, implying that the underlying mechanism is that of haploinsufficiency (Cortes-Lavaud et al., 2015, Kazenwadel et al., 2012). Indeed, using several *in vitro* techniques such as electrophoretic mobility shift (EMSA) and luciferase reporter assays, Hahn and colleagues showed that p.Thr354Met and p.Thr355del mutations reduce DNA binding affinity and transactivation abilities with potential dominant negative activities (Hahn et al., 2011, Chong et al., 2017).

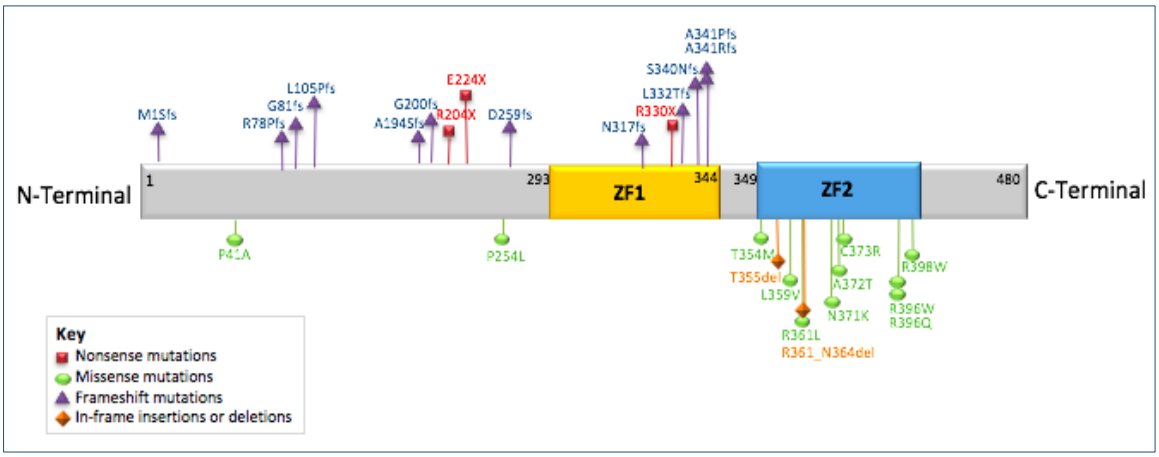


Figure 1.7 We "GATA" focus on GATA! Schematic representation of the GATA2 protein demonstrating the distribution of germline *GATA2* mutations in familial MDS/AML. Missense mutations (e.g. T354M) and deletions (e.g. T355del) in ZF2 are more associated with familial MDS/AML while nonsense and frameshift mutations in ZF1 are more associated with Emberger and MonoMAC syndromes. ZF1: zinc-finger 1; ZF2: zinc-finger 2. Figure adapted from (Nickels et al., 2013, Hyde and Liu, 2011).

Although the majority of *GATA2* germline mutations reside within the coding region, non-coding mutations in the conserved +9.5kb intron 5 enhancer element have been reported in patients with MonoMAC syndrome and in 10% of paediatric MDS patients, leading to decreased *GATA2* transcript levels and haploinsufficiency (Hsu et al., 2013). Also more recently, *GATA2* silent (synonymous) exonic mutations have been identified in patients with *GATA2* deficiency and were predicted to introduce splice sites defects leading to loss of mutant allele expression and premature RNA degradation (Kozyra et al., 2017, Wehr et al., 2018).

Such mutations constitute a new disease-causing mutational class in *GATA2* deficiency and necessitate further mechanistic studies in order to elucidate the functional impact of these germline non-coding variants.

1.2.3.4 Acquired Genetic Abnormalities Associated with *GATA2* Mutations

The mechanisms of clonal evolution in a *GATA2*-deficient background is not well understood. However, what became evident from the various studies so far was the high prevalence of trisomy 8 (15%) and monosomy 7 (40%) as recurrent cytogenetic aberrations linked with adverse outcomes (Bodor et al., 2012, Spinner et al., 2014, West et al., 2014b, Hirabayashi et al., 2017). In particular, Wlodarski and colleagues reported that ~72% of children and adolescents with MDS and monosomy 7 harbour germline *GATA2* mutations (Wlodarski et al., 2016). *GATA2* germline mutation screening should therefore be included in the work-up of all children and young adults with primary *de novo* MDS regardless of karyotype, family history or features of *GATA2* deficiency (Fisher et al., 2017). In addition, acquired mutations in the epigenetic regulator *ASXL1* have been demonstrated as an important second-hit in ~30% of germline *GATA2*-mutated patients and confer a poor prognosis (Bodor et al., 2012, West et al., 2014b, Micol and Abdel-Wahab, 2014, Wang et al., 2015). These acquired mutations in *ASXL1* are often accompanied by the presence of monosomy 7 acting as a diagnostic red flag although the order of these molecular events has not been elucidated.

Other secondary mutations encountered in *GATA2*-related disease are *EZH2*, *GATA1* and *HECW2,* reported in a *de novo* MDS patient with germline *GATA2* mutation (p.R330X) that was subjected to nonsense-mediated decay (NMD) (Fujiwara et al., 2014). Recurrent mutations in *STAG2* and *SETBP1* were also noted in a small subset of *GATA2*-deficient patients and are regarded as oncogenic drivers (Makishima et al., 2013, Inoue et al., 2015, Ding et al., 2017).

Altogether, these studies serve as another reminder for close monitoring of *GATA2*-mutated individuals for secondary cytogenetic or molecular genetic aberrations in order to trace the genetic evolution and progression of disease and identify patients in need for timely HSCT, considered to be the ideal choice of treatment modality for *GATA2*-mutated patients, with 4-year OS approaching 54% (Cuellar-Rodriguez et al., 2011, Spinner et al., 2014, Saida et al., 2016, Wlodarski et al., 2017).

Collectively, *GATA2* deficiency predisposes its carriers to familial MDS/AML and protean manifestations with cytopenia as its prominent clinical feature, highlighting the pleiotropic nature of this TF. A question remains in what determines the variable clinical presentation seen in some *GATA2*-mutated families despite carrying the same germline mutation. Thus far, there is no evidence of revertant somatic mosaicism in *GATA2* deficiency such as that encountered in *SAMD9L*-mutant AMLs and so the molecular mechanisms underpinning this enigmatic feature merits further investigation (Al Seraihi et al., 2018). I talk about that in more detail in section **1.3**.

1.2.4 How to Diagnose, Test and Manage Patients with Familial MDS/AML

Taken together, these germline predisposition syndromes are heterogeneous with regards to their causative genetic mutations, clinical manifestations and progression to overt MDS/AML. However, as a group, they all share the unique requirement for a high index of clinical suspicion to recognise the symptoms and perform appropriate genetic counseling, testing and mutation-specific clinical management (West et al., 2014a). These inherited forms of MDS/AML are more common than generally appreciated and the assumption that they are solely disorders of childhood no longer holds true, especially with the identification of germline *DDX41* mutations in patients aged over 62 years, making the recognition of these cases all the more difficult.

Therefore, accurate diagnosis and clinical management of affected individuals as well as asymptomatic carriers is of utmost importance and requires full knowledge and an in-depth understanding of the implications of inherited mutations, especially when considering suitable donors for allogeneic stem cell transplantation (Godley, 2014). Thankfully, we have come a long way since the discovery of germline *RUNX1* mutations in families with FPD/AML (Song et al., 1999) and are making significant progress in understanding the molecular pathogenesis of these series of diseases, where genetic testing is becoming more feasible in routine clinical practice and the candidacy of new genes is emerging from basic research and requiring functional validation.

1.2.4.1 Panel-based Molecular Testing and Germline Tissue Selection

Recently, genomic technologies have advanced to adopt multi-gene targeted resequencing panels for quick and reliable assessment of multiple genes in multiple samples simultaneously (with a coverage of ~1000x). This greater sequencing depth improves analytical sensitivity for VAF detection as low as 1% thus superseding traditional Sanger Sequencing with VAF cut-off of ~20% (Kadri et al., 2017).

The current diagnostic work-up in sporadic AML includes screening for mutations in *CEBPA*, *FLT3*, *NPM1*, and *RUNX1* however the compendium of genes is likely to expand following the identification of new mutations with prognostic and therapeutic impact or a role in familial disease (Dohner et al., 2017, Greenberg et al., 2017). In light of these findings, 12 out of the 14 familial MDS/AML disease loci (*ACD, ANKRD26, CEBPA, DDX41, ERCC6L2, ETV6, GATA2, MECOM, RUNX1, SRP72, TERC* and *TERT*) have been integrated into accredited diagnostic screening panels. In the UK, this is located at the West Midlands Regional Genetics Labs in Birmingham that the familial MDS/AML programme developed in collaboration with Dr Susanna Akiki

(https://ukgtn.nhs.uk). In the USA, a similar gene panel is available at the University of Chicago led by Dr Lucy Godley (https://dnatesting.uchicago.edu) while screening panels developed by other institutions can be found at GeneTests (www.genetests.org) and there is a global effort at present to increase access to genetic testing worldwide to help delineate population similarities and differences in mutational spectrum, prevalence and phenotype (Brown et al., 2017). Indeed, my ambition is to see a familial MDS/AML gene panel come into fruition in my home country or help establish one, soon.

Nevertheless, the aim of these initiatives is to ensure that every index case and their family members with suspected diagnosis of hereditary leukaemias can be genetically tested for mutations in these known genes and to capture the underlying genetic predisposition. Testing parental samples, if possible, might also be needed to confirm variant segregation with disease in the family (i.e. whether the germline variant is inherited or *de novo*). And clinicians are encouraged to integrate timely germline data into anticipatory patient care as not only does this information aid in diagnosis but it also has the potential to guide treatment decisions e.g. chemotherapy dosing and/or HSCT timing and donor selection (Guidugli et al., 2017).

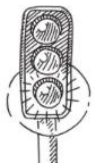
As many of the genes that confer familial inheritance can also be mutated as acquired secondary events or in sporadic MDS/AML (e.g. *RUNX1* and *CEBPA*), the use of germline DNA is absolutely essential to distinguish whether a mutation detected is of a somatic or germline origin. Obtaining germline material is however challenging since easily accessible samples such as PB, BM, saliva or buccal swabs are potentially contaminated with tumour DNA which can lead to false-positive results (Obrochta and Godley, 2018). The “gold standard” is constitutional DNA derived from cultured skin fibroblasts however it takes several weeks (4-6) to grow a sufficient number of cells and so this approach presents a time constraint that could complicate patient management

particularly when allogeneic HSCT is being considered (Bannon and DiNardo, 2016, Drazer et al., 2016). Therefore, in the absence of a matched germline sample, VAF assessment (e.g. if a mutation is present at heterozygous allelic frequency with VAF of >40%) and publically available datasets (e.g. the Exome Aggregation Consortium (ExAC) and TCGA) may provide valuable clues in support of a germline origin (Drazer et al., 2018). Also, follow-up testing of a remission DNA sample (where germline mutations would typically persist) or the presence of an unchanged VAF at multiple treatment time points may further support a germline origin of a variant (DiNardo et al., 2018). Where possible, it is advised to move towards an upfront paired tumour/remission testing (or tumour/germline testing at diagnosis) to maximise germline variant detection and identify acquired somatic mutations associated with disease progression. In fact, discussions are currently ongoing to see if this approach can become a standard diagnostic practice for every MDS/AML patient (Akpan et al., 2018, Tawana et al., 2018).

It is important to note however that these NGS-based gene panels will not tell the whole story as they are not equipped to detect large structural gene deletions, duplications or CNAs like the case of germline *RUNX1* deletions (Jongmans et al., 2010) nor non-coding germline mutations residing outside of the transcribed exons such as *GATA2* intronic variants (Hsu et al., 2013). This may lead to false-negative results and warrant referral for a more comprehensive research-based testing (e.g. by MLPA, CGH array or whole genome sequencing (WGS)), especially when the familial presentation is well established clinically. While non-coding variants may predispose to leukaemia, the interpretation of this data is less clear at the present time and may require functional validation. Moreover, it is not just testing for mutations as other phenotypic assays such as telomere length measurement can be employed for the diagnosis of patients with IBMFS by means of flow-fluorescence *in situ* hybridisation (FISH) (Walne et al., 2013a).

1.2.4.2 New Gene Discovery and Variant Assessment

> *"As we know, there are **known knowns**; these are things we know we know. We also know there are **known unknowns**; that is to say we know there are some things we do not know. But there are also **unknown unknowns** – the things we don't know we don't know….it is the latter category that tend to be the difficult one."*
>
> *– Donald Rumsfeld - US Secretary of Defence*



Remarkably, germline mutations in the aforementioned genes represent merely 60% of familial cases investigated to date and so approximately 40% of families remain without an underlying genetic aetiology, where variants in the known genes were not identified, at least from interrogation of the coding genome. Therefore, ongoing investigations currently being carried out in my host laboratory and others are focused on identifying novel inherited disease lesions in candidate genes and expanding the current germline mutational repertoire (Rio-Machin et al., 2018a). Briefly, this is done by integration of WES data from these uncharacterised families (where two or more members are diagnosed with a haematological malignancy (AML, MDS or BM failure) including at least one member with MDS and/or AML in each family) and filtering genes based on stringent criteria (see box - next page). This is then followed by variant confirmation by Sanger sequencing and functional validation of select candidates *in vitro* or *in vivo* (e.g. by shRNA gene knockdown or CRISPR-Cas9 genome editing) to determine if these putative candidates are *bona fide* disease-causing in familial MDS/AML. This is exemplified by the recent discovery of novel germline variants in *CHEK1, MECOM*, *SAMD9* and *SAMD9L* segregating in families with MDS/AML and/or other accessory phenotypes (e.g. MIRAGE syndrome) (Wartiovaara-Kautto et al., 2018, Ripperger et al., 2018, Narumi et al., 2016, Chen et al., 2016).

Given the increasing use of NGS technologies both in research and clinical setting, not only do these tour de force tools offer an exceptional opportunity for comprehensive and unbiased gene hunting but also an immense challenge as to how one can systematically interpret the data and assign true pathogenicity when over 500 non-synonymous variants are detected per sample. Therefore, using the features of known germline disease-causing variants (known knowns) as a guide, our laboratory has established the following criteria of filtering WES data to narrow down the search and exclude non-pathogenic variants or frequent single nucleotide polymorphisms (SNPs) taking into account:

1) Whether a variant has been previously reported in the literature or mutated in sporadic disease and assessing variant frequency in healthy populations using publicly available databases such as (ExAC, HAPMAP, dbSNP, ClinVar or gnomAD) where for example ExAC variant frequency of <0.0001 is considered novel;

2) Variant type (missense or protein-truncating etc.) and predicting the biochemical function of variants by virtue of *in-silico* tools such as (PolyPhen2, MutationTaster, SIFT, Provean, GERP, cBioPortal, OMIM or CADD) where for example variants with PolyPhen2 scores >0.850 are predicted to be damaging and most importantly;

3) Pedigree information and whether the variant is present in at least two or more families (with at least two members analysed per family) can be a useful source to determine if there is a segregation of the variant with disease (this is not often the case unfortunately as in some families, only the index case is analysed, owing to limited sample availability).

Based on these population, computational and segregation data, variants can be classified into different categories varying from benign to pathogenic (or actionable) and variants of uncertain significance (VUS) (Richards et al., 2015, Baudhuin et al., 2016). However, despite the availability of these *in silico* prediction tools, interpreting with certainty the pathogenicity and functional consequences of the detected variants can still be very difficult especially when the variant corresponds to the latter category. Additional evidence (e.g. clinical phenotype and/or family

history) would therefore be needed to support variants' association with disease (Richards et al., 2015). For example, a family member presenting with lymphoedema, MDS and a novel *GATA2* VUS (known unknown) would raise a suspicion of a *GATA2* deficiency syndrome and prompt variant re-classification to likely pathogenic (Brown et al., 2017). The difficulty arises when a novel VUS is identified in a family without a clear candidate gene (unknown unknown). Therefore, knowing the function of the encoded protein, its impact on leukaemogenesis (if any) and fine-tuning the filtering criteria (e.g. by increasing the number of families analysed or tweaking ExAC frequency cutoffs) would help shortlist credible candidates and shine light on these needles lurking in the genomic haystack.

Given the rarity of families with MDS/AML at present and the marked genetic and clinical heterogeneity, these investigations require a network of collaboration to consolidate these findings, which our group and several others around the world (e.g. Hamish Scott's group from Adelaide, Australia, Marcin Wlodarski's group from Freiburg, Germany and Lucy Godley's group from Chicago, the United States) have begun to establish over the past few years.

1.2.4.3 Considerations for Clinical Management and Genetic Counselling

The identification of individuals with germline predisposition to myeloid malignancies for referral for genetic counselling and testing remains challenging for several reasons: there is a lack of awareness regarding the existence of these predisposition syndromes, difficulty in obtaining germline material coupled with lack of widespread availability of testing and incomplete family/personal history (DiNardo et al., 2018). Clinical vigilance of disease symptoms underpins the detection of these cases. Fortunately, since the start of my PhD in 2015, several disease genes have been identified including *DDX41* (Polprasert et al., 2015), *SAMD9* (Narumi et al., 2016) and *SAMD9L* (Chen et al., 2016, Tesi et al., 2017) and current screening assays should

incorporate these new findings. While these discoveries brought a momentum to the field of familial leukaemia research, it is likely that they merely represent 'the tip of the iceberg' and so routine testing of germline or remission DNA is essential to identify more familial cases.

Currently, there is little practical information or guidelines on how diagnostic evaluation should be performed in these cases, particularly in the context of clinical trials, and recommendations for surveillance often depend on the nature of the predisposing syndrome and are emerging from several research groups in the field. **Figure 1.8** depicts an algorithm adapted from expert opinion for diagnostic work-up and to identify appropriate patients for referral to the clinic (Churpek et al., 2013, Drazer et al., 2016, Akpan et al., 2018). Generally, it is recommended to refer patients who have a family history (2 or more members within at least 2 generations) of haematological malignancies or a personal history of multiple cancers or any features suggestive of a haematological malignancy such as unexplained cytopenias, severe anaemia or bleeding. The occurrence of MDS in a younger patient should also raise suspicion of a germline syndrome (e.g. *GATA2* deficiency) (Brown et al., 2017). Close clinical follow-up and monitoring of affected as well as asymptomatic carriers, where possible, is critical as they are advised to undergo a baseline bone marrow biopsy and cytogenetic analysis to asses for occult malignancy in addition to twice annual physical examination and complete blood count (CBC) (Drazer et al., 2018).

Patients under consideration for allogeneic HSCT with an HLA-matched first-degree relative donor warrant expedited germline genetic testing to rule out a familial MDS/AML syndrome and exclude asymptomatic mutation carriers, who may have normal CBCs, from the transplant scenario (Nickels et al., 2013). Indeed, reports on engraftment failure and donor-derived leukaemia (DDL) act as an important reminder of the potential consequences when germline

mutations are not considered and tested first hand. Examples of DDL cases include those harbouring mutations in *CEBPA* (Xiao et al., 2011), *DDX41* (Berger et al., 2017, Kobayashi et al., 2017) and *GATA2* (Galera et al., 2018). However, in the absence of a known germline mutation but a strong family medical history, the use of a matched unrelated donor is advisable (Niemeyer and Mecucci, 2017).

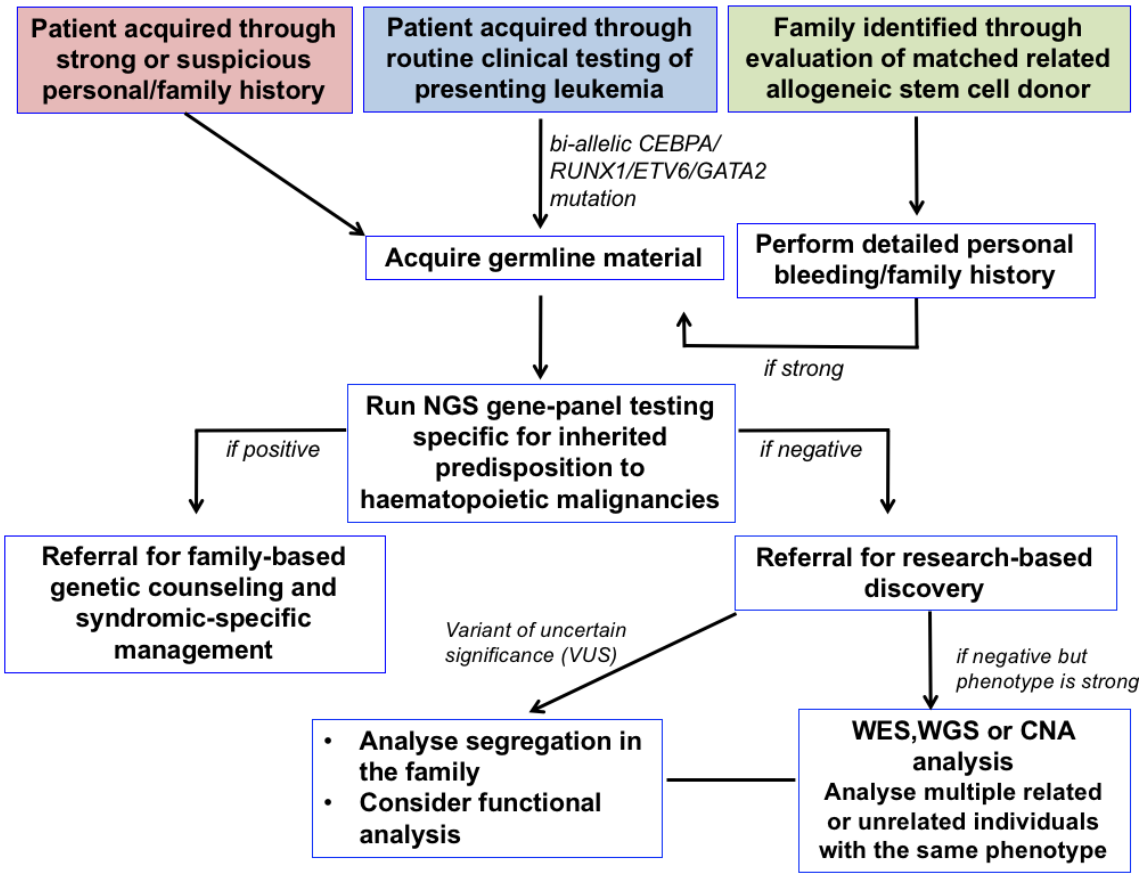


Figure 1.8 An algorithm for work-up of patients with familial predisposition of myeloid malignancies. Figure adapted from (Drazer et al., 2016).

Finally, prompt genetic counselling remains an integral component for the optimal care of these patients and their families, to help them better understand the principles of human genetics as well as overcome the impact of these inherited disorders on their emotional well-being. Genetic counsellors also help patients and family members become aware of the benefits and potential risks of genetic testing, possible test outcomes and implications for treatment and donor

selection and above all, respect their choice, privacy and autonomy. This information forms the basis of the informed consent process for testing and disclosure of results (Nickels et al., 2013). Affected couples may seek reproductive counseling for prenatal testing options including *in vitro* fertilisation (IVF) and preimplantation genetic diagnosis (PGD) to determine disease risk in their offspring and/or select embryos lacking the germline defect (Niemeyer and Mecucci, 2017). Patients are also encouraged to participate in research to allow for the discovery of new variants and susceptibility syndromes (Brown et al., 2017). Indeed, patients and their families feel empowered knowing their contributions and stories can have a major impact on the research and medical field by helping other families going through similar experiences they have been through.

1.2.4.4 Unanswered Questions in Familial Leukaemia Research

The field of familial leukaemia research is evolving rapidly. Much remains to be learned considering the rarity of these familial cases and as with all good studies, there are more questions than answers. For example, what is the true prevalence of familial MDS/AML syndromes in adult populations (currently amounts to <5% of all MDS/AML cases) now that more susceptibility genes are likely to be identified? Does the scientific community need a uniform and accessible algorithm for research-based gene discovery and variant triage? Secondly, should all patients with MDS/AML undergo germline testing at the time of diagnosis and irrespective of family history, or be included in a tumour/germline panel-based testing pipeline as part of a standard procedure in routine clinical practice? And thirdly (and somewhat controversial), considering that germline mutations in familial leukaemia can display reduced penetrance, should prophylactic allogeneic HSCT be offered to asymptomatic mutation carriers or indeed to those presenting with symptomatic manifestations but have not yet developed full-blown malignancy? And how can we identify these unaffected carriers in the first place?

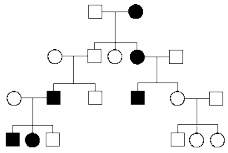
(Drazer et al., 2016, Akpan et al., 2018, Rio-Machin et al., 2018b, Tawana et al., 2018).

Addressing these questions and more as they develop over the coming years will likely require the collaborative efforts of an international clinical and research community who can provide consensus and practical guidelines. All this information will hopefully lay a strong enough foundation for optimal management and care of this unique patient population, with novel molecular insights that may ultimately lead to the implementation of new diagnostic measures and mechanism-based therapeutic interventions for both familial and sporadic forms of the disease.

1.3 Penetrance of Germline Mutations

*"They f*** you up, mum and dad.*
They may not mean to, but they do.
They fill you with the faults they had
And add some extra, just for you"

– Phillip Larkin, This Be The Verse



"They tuck you up, mum and dad.
They read you Peter Rabbit, too.
They give you all the treats they had
And add some extra, just for you"

– Adrian Mitchell, This Be The Worst

One of the conundrums in familial cancers in general and leukaemias in particular is the notion that not every individual carrying a particular variant (or genotype) will eventually exhibit clinical symptoms (or phenotypes) associated with the disease in question – a phenomenon known as **"incomplete or reduced penetrance"** (Cooper et al., 2013). Throughout this thesis, the term "reduced penetrance" is used to describe both penetrance (proportion of mutation carriers showing a phenotype – not always 100%) and variable expressivity (severity of phenotype) although I personally like to call it "mutation shyness" as the same *bona fida* disease-causing mutation may not be expressed equally in all individuals who carry it.

This inter-individual variability poses a great challenge for understanding how genetic variants manifest in disease (or why certain carriers do or do not manifest it) and a caveat for the prognosis of patients' disease outcomes based solely on their genetic information (Castel et al., 2018).

Reduced penetrance is more evident in autosomal dominant disorders when simple rules of Mendelian genetics are not followed. These include muscular dystrophy (caused by *LMNA* mutations), retinoblastoma (*RB1* mutations), Huntington's disease (*HTT* mutations) and breast and ovarian cancers (*BRCA1* and *BRCA2* mutations) (Vytopil et al., 2002, Lefevre et al., 2002, McNeil et al., 1997, Chen and Parmigiani, 2007). However, the precise molecular explanation of such occurrence is not well understood.

1.3.1 Reduced Penetrance in Inherited *GATA2*-mutated MDS/AML

Scientific investigations of familial leukaemia have not addressed the question of reduced penetrance in earnest which could have important connotations for the counselling and management of patients and their families. Germline mutations that display low penetrance tend to be associated with a higher frequency of asymptomatic "silent" carriers such that the clinical manifestation may not be evident in one generation but can still be transmitted, through unaffected parents, to the next generation where it can once again manifest itself – so-called generation skipping (Cooper et al., 2013). We have noted that this is indeed a feature within certain *GATA2*-mutated families, particularly those carrying germline missense LoF mutations and seems less likely in other types of *GATA2* mutations, suggesting that penetrance is potentially a mutation-dependent occurrence. It is pertinent to note that penetrance may vary not only with mutation type, but also with the location of the mutation in the protein. An example is germline *CEBPA* mutations in the C-terminal domain being more apt to exhibit

reduced penetrance as opposed to N-terminal mutations where penetrance is almost near complete (Tawana et al., 2015, Pathak et al., 2016).

In the initial study reported by Hahn and colleagues from Australia describing three *GATA2*-mutated MDS/AML pedigrees spanning multiple generations, all affected members who were tested in those families (13/19) carried the heterozygous missense mutation (p.Thr354Met). However, not all family members (6/19) harbouring this germline mutation developed symptoms of haematological malignancy (e.g. cytopenia or neutropenia), remaining symptom-free well into their 50's or 60's (**Figure 1.9** – pedigrees 1, 2 and 3) (Hahn et al., 2011). Since this observation, few additional families where disease penetrance varied from 50 to 70% came to light concomitantly (**Table 1.5**), including a five-generation *GATA2*-mutated MDS/AML family published by our group carrying p.Thr354Met and monosomy 7 (Bodor et al., 2012) (**Figure 1.9** – pedigree 4) (discussed in detail in **Chapter 3**).

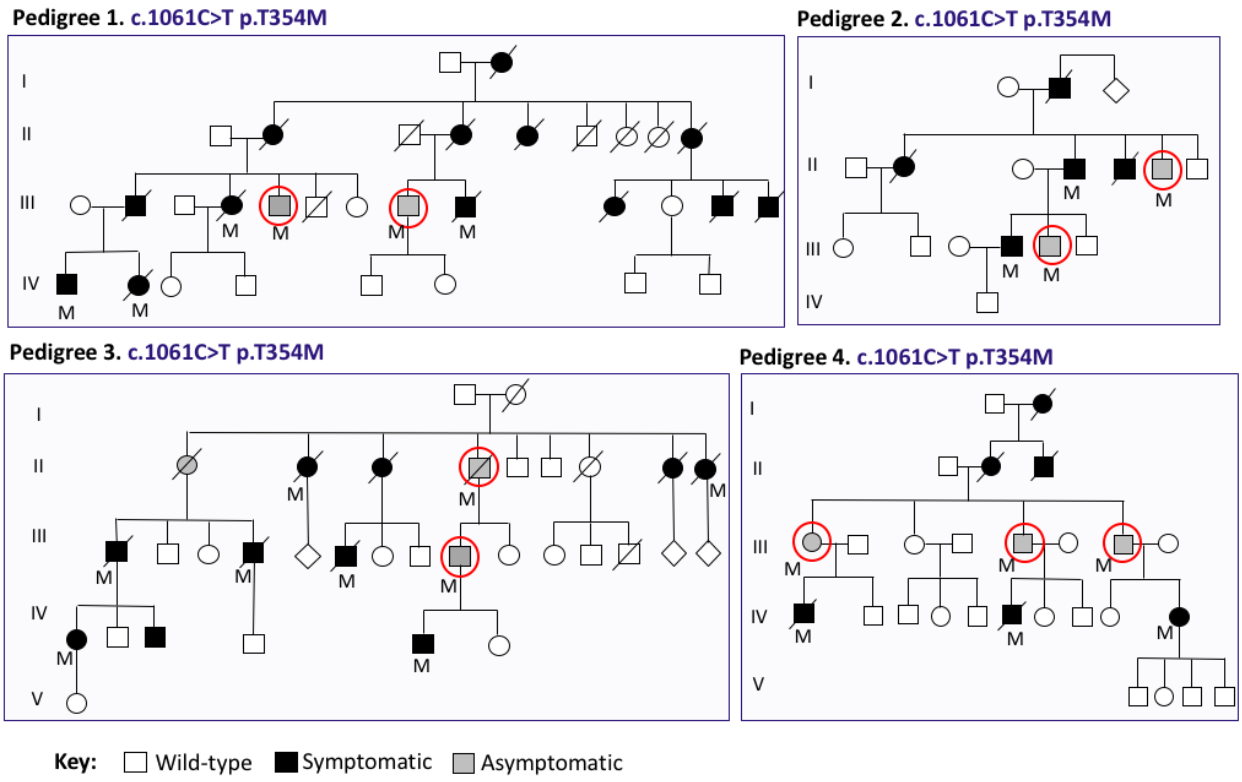


Figure 1.9 *GATA2* (p.T354M)-mutated MDS/AML families with reduced penetrance. Pedigrees re-drawn from (Hahn et al., 2011) and (Bodor et al., 2012). Asymptomatic carriers are circled in red. M: family member carrying the p.T354M mutation.

Why do we care about a phenomenon that is a rare exception to the norm? True estimates of reduced penetrance are masked by limited genetic and phenotypic screening of extended family members and absence of comprehensive family medical histories, suggesting it may be more common than previously considered (Shawky, 2014). Measuring penetrance quantitatively, however, has been a subject of controversy as what constitutes a "symptomatic", "asymptomatic", "affected" or "obligate" carrier may vary from one institution to another and therefore, a consensus definition of these nomenclature, at least within the realm of familial AML, is needed. Ultimately, reduced penetrance represents an added complication for clinicians and genetic counsellors alike, particularly when predicting disease risk in future generations (Drazer et al., 2016). Reduced penetrance is also likely to present an impediment to the interpretation of NGS data (e.g. based on ExAC allele frequencies) and so it must be taken into account when classifying the pathological significance of variants (Plon et al., 2008, Richards et al., 2015). Therefore, careful family clinical evaluation and identification of asymptomatic 'silent' mutation carriers, to include them on surveillance and genetic testing but critically to exclude them as potential stem cell transplant donors, is of great importance. It is hoped that deciphering the molecular mechanisms governing this enigmatic feature of human inherited disease and uncovering potential genotype-phenotype correlations will help improve our understanding of leukaemia pathogenesis and aid in genetic counselling of at-risk germline mutation carriers.

Aside from familial studies, estimates of penetrance can also be obtained through population-based linkage analyses and genome wide association studies (GWAS) using large disease cohort datasets. For example, several studies showed that certain common variants *in cis* have the potential to modify the penetrance of *BRCA1*- and *BRCA2*-associated breast cancers (Gaudet et al., 2010, Couch et al., 2013).

Another example is a recent study by Tuuli Lappalainen's lab who used population-scale functional genomics to assess how *cis*-regulatory modifiers of penetrance affect disease risk in individuals with cancer and autism spectrum disorder (ASD) (Castel et al., 2018).

1.3.2 Molecular Factors Associated with Reduced Penetrance

What accounts for the clinical variability between symptomatic carriers sharing an identical germline mutation or determines progression to overt malignancy in asymptomatic ones remains unclear. It is difficult to pinpoint a single mechanism or answer but it is likely to be down to a combination of complex genetic and environmental factors. These include: mutation type, differential gene or allelic expression, *cis*- or *trans*-regulatory allelic variations, epigenetic modifications, copy number alterations, age or simply, luck.

The premise is that genotype is not the sole determinant of phenotype but rather the intersection of genetic variants, environment, chance and other genetic modifier effects (Mukherjee, 2016). Identification of such molecular events can blur the boundaries between simple monogenic and complex genetic diseases, challenging the traditional dogma that a human disease can be caused by a "single" mutated gene (Badano and Katsanis, 2002). Concurring with these views and in the context of familial leukaemia, one would argue that germline mutations with reduced penetrance (as is the case for *GATA2*) may contribute to disease pathogenesis but are perhaps insufficient on their own to trigger the development of overt malignancy and in so doing, require the presence of additional co-operating events. Whilst each of these factors may influence disease penetrance in their own right in some fashion, different combinations are likely to exert different effects on the severity of symptoms and so for the purpose of this thesis, the involvement of these molecular factors in disease penetrance shall be tested.

1.4 Aims & Objectives

"Find out the cause of this effect,
Or rather say, the cause of this defect,
For this effect defective comes by cause."

– William Shakespeare, Hamlet

Over the past two decades, many research studies have provided examples of inherited forms of myeloid malignancies and a glimpse of their genetic and phenotypic complexities, with germline mutations reported in ~14 disease genes so far, leading to their incorporation as a separate diagnostic entity into the 2016 WHO classification system. However, a number of questions are yet to be answered in this emerging field, including disease penetrance and what accounts for the clinical heterogeneity seen within a given family sharing the same germline mutation. Therefore, in light of these observations in *GATA2* p.Thr354Met MDS/AML pedigrees and the availability of sequential samples and NGS/molecular tools at one's disposal, the primary aim of the study addressed in my thesis was to:

Investigate the molecular mechanisms underlying reduced penetrance of germline *GATA2* mutations. In particular, this work aimed to understand the molecular basis of disease initiation and identify molecular features that distinguish symptomatic from that of asymptomatic family members by:

- Employing in-depth molecular profiling – with the hypothesis that penetrance might reflect on the acquisition of secondary somatic mutations (or second hits) that differentiates between these two groups of mutation carriers and;
- Investigating *GATA2* expression levels – with the hypothesis that penetrance may be modulated by changes in *GATA2* expression (and all its elements: global, regulatory and allelic dosage) defining symptomatic and asymptomatic carriers and potentially paving the mutational path towards malignancy.
- Elucidating the molecular mechanisms by which GATA2 is regulated in these *GATA2*-mutated patients.

Chapter 2. Materials & Methods

2. Materials and Methods

"Progress in science depends on new techniques, new discoveries and new ideas, probably in that order."

– Sydney Brenner

2.1 Familial MDS/AML patient sample collection and study approval

Bone marrow (BM), peripheral blood (PB) or salivary material were obtained from select members of the *GATA2*-mutated MDS/AML family investigated in this thesis (described in detail in **Chapter 3** with a summary of the clinical history in section **3.2**). These include: asymptomatic mutation carriers (III.5 and III.7), wild-type individuals (IV.8 and IV.9) and multiple time-points from the symptomatic patient (IV.10: y.1, 3, 4 and 6) obtained through routine clinical follow-up. Tumour DNA samples were extracted from BMs of (IV.1 and IV.6) deceased MDS/AML cousins at diagnosis while constitutional or germline DNA was sampled from (IV.6) during disease remission. Samples were collected via referrals from clinicians managing these patients at St. Bartholomew's hospital, London. Informed, written consent was obtained for use of samples for research with ethical approval (06/Q0401/31) received in accordance with the Declaration of Helsinki.

2.2 DNA mutational profiling and sequencing

PCR amplification followed by Sanger sequencing of regions encompassing variants of interest was performed. Briefly, primers designed to anneal specific sequences surrounding a particular variant were used for target DNA amplification (see **Table 2.1** for primer sequences and annealing temperatures (T_A)). PCR amplicons were run on an agarose gel and cleaned to remove excess reagents and primer dimers. Purified PCR amplicons were then sequenced on forward and reverse directions via fluorescent cycle sequencing. Sanger sequencing products were analysed on the Life Technology 3730xl DNA Analyzer and trace chromatograms visualised on the BioEdit

software and compared against a reference human genome.

2.2.1 Genomic DNA extraction and quantification

Genomic DNA was extracted from patient PB or BM aspirates using the DNeasy Blood & Tissue Kit (Qiagen) following standard procedures. For samples recovered from the liquid nitrogen, cells were thawed at 37^{o}C water bath and transferred to a fresh medium. Approximately 5 million cells were centrifuged (300g for 5 minutes at 18^{o}C) and resuspended in 200µl Dulbecco's Phosphate-Buffered Saline (PBS) to remove any dimethyl sulfoxide (DMSO) residues. 20µl of Proteinase K and 200µl of lysis buffer (AL) were next added to the cells and the mixture vortexed prior to incubation at 56^{o}C for 10 minutes. The later addition of 200µl of 100% ethanol allowed DNA to precipitate which was then transferred onto silica-based membranes in DNeasy Mini spin columns and centrifuged at 6000g for 1 minute. DNA bound to the membrane columns then underwent sequential washes using ethanol-containing buffers: 500µl of buffer AW1 (6000g for 1 minute) followed by 500µL of buffer AW2 (20,000g for 3 minutes). Finally, purified DNA samples were eluted in 100µl of elution buffer (EB) or nuclease-free water prior to long-term storage at -20^{o}C.

Nucleic acid concentration and quality were measured (in ng/µl) following extraction using the NanoDrop^{TM} ND-1000 Spectrophotometer (Thermo Scientific, Wilmington, MA) to establish the 260nm/280nm and 260nm/230nm absorbance ratios, where ratios between 1.8 and 2.0 were regarded as "pure" DNA and/or RNA. Any ratios significantly lower than these values potentially indicate sample contamination (e.g. by residual phenols or proteins). Alternatively, a flurometric-based approach (Qubit® 3.0) and the dsDNA Broad Range Assay Kit (Life Technologies) was used for DNA quantification due to its higher sensitivity and specificity compared to its absorbance-based counterpart.

2.2.2 Direct polymerase chain reaction (PCR)

Direct PCR was used to amplify precise regions of the DNA. PCR reactions were performed using the 2x ReddyMix™ PCR Master Mix (Life Technologies) that contains a Thermoprime *Taq* DNA Polymerase and a *Taq* buffer with a red loading dye to aid in visualisation of PCR products. Primers sequences encompassing variants of interest (coding or non-coding regulatory regions) were designed using the IDT PrimerQuest tool: (http://www.idtdna.com/primerquest/home/index) and sequences are listed in **Table 2.1**. Each 10µl reaction contained 4µl of ReddyMix, 4µl of water, 0.5µl of forward and reverse primers (10µM concentration per primer) and 1µl of DNA (or cDNA) template (~50ng/µl). A non-template control (NTC) (1µl of water) was also added in each experiment to detect possible reagent contamination and/or non-specific amplification.

The PCR cycling conditions consisted of an initial denaturation step (at 95°C for 10 minutes), followed by 35 cycles of DNA denaturation (at 95°C for 30 seconds), primer annealing (at 50-65°C for 30 seconds) and extension (at 72°C for 60 seconds) followed by a final extension step (at 72°C for 10 minutes). The optimal annealing temperature (T_A) for each primer pair/PCR reaction (**Table 2.1**) depends on multiple variables (e.g. primer nucleotide length (20-30) and GC content (40-60%)) and is ideally 5°C below the lowest primer's melting temperature (T_m) (New England BioLabs (NEB)).

Table 2.1 Primer sequences used in this study together with their annealing temperatures. These primers were designed using different programs including: Primer3plus, Primer-BLAST, IDT PrimerQuest Tool and MethPrimer. The vector M13 insert primer sequence was obtained from the TA-cloning protocol (Invitrogen).

Gene	Forward Primer	Reverse Primer	Annealing Temperature (T_A)
Intronic PCR primers for Genomic DNA amplification			
GATA2 Exon 4	GACTCCCTCCCGAGAACTTG	GCGTCTGCATTTGAAGGAGT	58°C
GATA2 Exon 5	TTAGCCCTCCTTGACTGAGC	AGCCAAGCTGGATATTGTGG	58°C
GATA2 Exon 6	GTTGCTGGAGGAAGGAACTG	AACTGTCCATGCAGGAAACC	58°C
ASXL1	GGTTAAAGGTCAGCCCACTTA	CAGTAGTTGTGTTCGCTGTAGA	56°C
Exonic RT-qPCR primers for cDNA amplification			
GATA2 Exon 5	ACTCATCAAGCCCAAGCGAA	CTTCATGGTCAGTGGCCTGT	60°C
GAPDH	CCATCACCATCTTCCAGGAG	GAGATGATGACCCTTTTGGC	56°C
Overlapping PCR primer sets covering GATA2 regulatory regions			
Promoter 1 (1)	GGAAGGTGGAAGTGGAAGAAA	GCTGCTCAAGTCTGTCTATT	54°C
Promoter 1 (2)	GGCAGGCAATAGACAGACTT	GGGCGTCTGTGTACCATTAT	57°C
Promoter 1 (3)	AGCTTGGGACACGTCTTTAC	GACTCCCACACTCACTTTCT	57°C
Promoter 2 (1)	CGCCAGATACACATACTGATCTC	GCTGGCTTGGGCTTCTTA	57°C
Promoter 2 (2)	GACTCCTGCACAGACATGAA	GGGCAGTTGGTGGTTAGTTA	57°C
Promoter 2 (3)	TCCGCAATTCCCGAACC	GGGCAGTTGGTGGTTAGTTA	57°C
Promoter 2 (4)	GGCCTCCCTAGCAGTAACTAA	AGGTGACTTAGAAGACGGAGAC	58°C
Enhancer 1	TTTGCAGAGTGGAGGGTATTAG	ATGGAGTCACCTATACTGTGTATTT	54°C
Enhancer 2 (1)	CCTCCCAAAGTGCTGAGATTAC	CTGGGAGTTGGAAGTTGTAGTG	58°C
Enhancer 2 (2)	CTCCTAGGTTCATGCCATTCTC	TGGTAAGGCTATGAGGGATACA	57°C
Enhancer 2 (3)	CCCAGGTTCAAGCGATTCT	GGCTGAGGGTATCAACAAGAG	58°C
Enhancer 2 (4)	TTTCTCAGGCCTCCTGATTTC	CCTGCTGACCATGAGTGATT	57°C
Enhancer 2 (5)	GCCTAAGAGCGTCTTGCTAAA	CTCGTGAGAGTGGTCTGAATTG	57°C
Methylation-specific PCR (MSP) GATA2 primers			
Promoter 1 (Methylated)	CGGGTATTTTTTTGTTTTTTGC	TAACCTCGCTACCTTCCTAACG	51°C
Promoter 1 (Unmethylated)	TTTTGGGTATTTTTTTGTTTTTTGT	CTAACCTCACTACCTTCCTAACACT	50°C
Promoter 2 (Methylated)	TTCGTTTTATGTTTGTGTAGGAGTC	AACTAAAAATAAAATACGCCTCGAC	52°C
Promoter 2 (Unmethylated)	TGTTTTATGTTTGTGTAGGAGTTGG	AACTAAAAATAAAATACACCTCAAC	50°C
Allele-specific bisulphite-specific PCR (BSP) GATA2 primers encompassing P2 SNP [C/A]			
Promoter 2 (Methylated)	GGTTTTTGAGAGTGAAGGAGTTTC	CTATACAAAAATCGACAACTAACGC	53°C
Promoter 2 (Unmethylated)	GGTTTTTGAGAGTGAAGGAGTTTT	TATACAAAAATCAACAACTAACACC	50°C
TA Vector M13	GTAAAACGACGGCCAGT	CAGGAAACAGCTATGAC	50°C
Allele-specific ChIP GATA2 primers (shorter amplicon sizes for ChIP-enriched DNA)			
Promoter 2 (141bp)	AACCCCAAACTTACACACGC	CAGCTCCTACCCTGTAAGCC	59°C
Promoter 2 (95bp)	GAGAGTGAAGGAGTTCCGGC	CCCCAGCTCCTACCCTGTAA	61°C

2.2.3 Agarose gel electrophoresis and purification of PCR products

Agarose gel electrophoresis was used to confirm the size and presence of the amplified PCR products. 1.5% (w/v) agarose gels were prepared using 2.25g of agarose with 150mL of 1x Tris-Borate-EDTA (TBE) buffer (10x TBE contains: 108.9g Tris base, 55.7g boric acid and 7.4g EDTA in 1 litre of H2O). The agarose was melted by heating for approximately 4 minutes in a microwave and once partly cooled down, 7.5µl of 1x GelGreen™ (Biotium, Hayward, CA) was added. The gel was then poured into a cassette with fitted combs and allowed to set. PCR products along with a 5µl of 1kB DNA ladder (exACTGene™, Fisher Scientific) were loaded into separate wells and gel electrophoresis was performed at 100V (for ~60 minutes) in a 1xTBE running buffer and subsequently visualised under ultra violet (UV) light to confirm the size and presence of PCR products.

The correct-sized PCR bands were then excised from the agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's protocol. Alternatively, PCR products were enzymatically cleaned from superfluous primers or unincorporated dNTPs using Exonuclease I (Exo I) and shrimp alkaline phosphatase (SAP) (USB, Cleveland, Ohio, U.S.A). The Exo-SAP stock solution constituted of 1µL exonuclease, 20µL SAP and 179µL water. 4µL of Exo-SAP was added to 5µL of PCR products and the mixture underwent thermal cycling at 37^{o}C for 15 minutes, 80^{o}C for 15 minutes and 10^{o}C for 10 minutes before proceeding to the sequencing reaction.

2.2.4 Sanger Sequencing

"Of the three main activities involved in scientific research, thinking, talking and doing, I much prefer the last and am probably best at it…"

– Frederick Sanger

Purified PCR products were sequenced in both forward and reverse directions using the BigDye™ Terminator Cycle Sequencing kit version 3.1 (Applied Biosystems). Briefly, the 10µl sequencing reaction consisted of 0.25µl Big Dye Terminator mix™, 1.875µl of 5x sequencing buffer, 0.5µl of forward or reverse primers (final concentration is 10µM per primer) (please see **Table 2.1** for primer sequences used in this study), 1µl of the purified PCR product and 6.375µl of RNase-free water. The hybridisation reaction entailed an initial denaturation step at 96^{o} C for 1 minute followed by 26 cycles of denaturation at 96^{o}C for 30 seconds, primer annealing at 50^{o}C for 15 seconds and extension at 60^{o}C for 4 minutes. Purified PCR products/primer mixture were then diluted accordingly and shipped to GATC Biotech Sanger sequencing services (Eurofins) in Germany for clean-up and DNA size separation via bidirectional capillary electrophoresis using the ABI PRISM 3730xl Genetic Analyser (Applied Biosystems, CA).

2.2.5 Identification and validation of mutations

Sanger sequencing trace chromatograms were visualised and interpreted using the BioEdit software (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and compared against a human reference genome sequence obtained from the Ensembl genome browser (https://www.ensembl.org/index.html) to validate the sequenced variants. All variants were confirmed by at least three independent PCR and sequencing experiments.

2.2.6 TA-cloning of PCR products

To define the allelic ratios of mutations or quantify mutant vs. WT alleles, PCR products were cloned into TA-plasmid vectors which were then transformed into chemically competent E. coli cells and subsequently sequenced to obtain individual clones (each clone corresponding to one allele). This was achieved using the TA cloning™ kit (Invitrogen).

Briefly, PCR products were generated using ReddiyMix master mix, as described in section **2.2.2**. The single 3'(A)-overhangs on PCR amplicons can attach to 3'(T)-overhangs on the linearised pCR2.1® vector, thereby facilitating the ligation. The 10µl ligation reaction consisted of: 2µl of pCR2.1® vector (50ng), 1µl of ExpressLink T4 DNA ligase, 1µl of ligation buffer (10x), 5µl of nuclease-free water and 1µl of the PCR product (~10ng insert DNA). This reaction mixture was then incubated for at least 1 hour at room temperature.

The cloned PCR products were then transformed into chemically competent E. coli (One Shot™ TOP10 cells, Invitrogen). Approximately 2µl of ligation mixture was added to each thawed vial of E. coli cells (50µl) and the mixture was left on ice for 30 minutes. Cells were then heat-shocked for 30 seconds at 42°C, to enable vectors to enter the bacterial cells, before transferring them back on ice for further 2 minutes. 250µl of SOC medium (2% tryptone, 0.5% yeast extract, 10mM NaCl, 2.5mM KCl, 10mM MgCl2, 10mM MgSO4, 20mM glucose) was next added to each vial of transformed cells and left incubated at 37°C for 1 hour on an incubator shaker at 225rpm. Later, approximately 100-150µl of the transformed bacterial cells/vials were spread onto Luria Bertani (LB) agar plates containing an antibiotic e.g. ampicillin (100µg/ml) and layered with 40µl of X-Gal solution (Bioline) (40mg/ml) diluted in dimethylformamide (DMF), to facilitate blue/white selection of colonies.

Lastly, agar plates were incubated overnight at 37°C to allow for colony formation and were transferred the next day to 4°C for 2-3 hours to allow for proper colour development.

Following overnight incubation, in the bacterial colonies with no insert DNA, the empty vector promotes the production of β-galactosidase, encoded by the *LacZ* gene, which is hydrolysed by X-Gal thereby producing a blue colour (these were discarded). The selection of white colonies, however, implies that the vector and insert are both intact within the bacterial clones, therefore the formation of an active β-galactosidase is disrupted (**Figure 2.1**). Following an overnight incubation at 37°C and shaking of ~25 white colonies (per plate) seeded separately into LB broth containing ampicillin, plasmid DNA purification was performed using the QIAprep Spin Miniprep Kit (Qiagen) as per the manufacturer's instructions followed by Sanger sequencing of single clones using the vector M13 primer pair (**Table 2.1**).

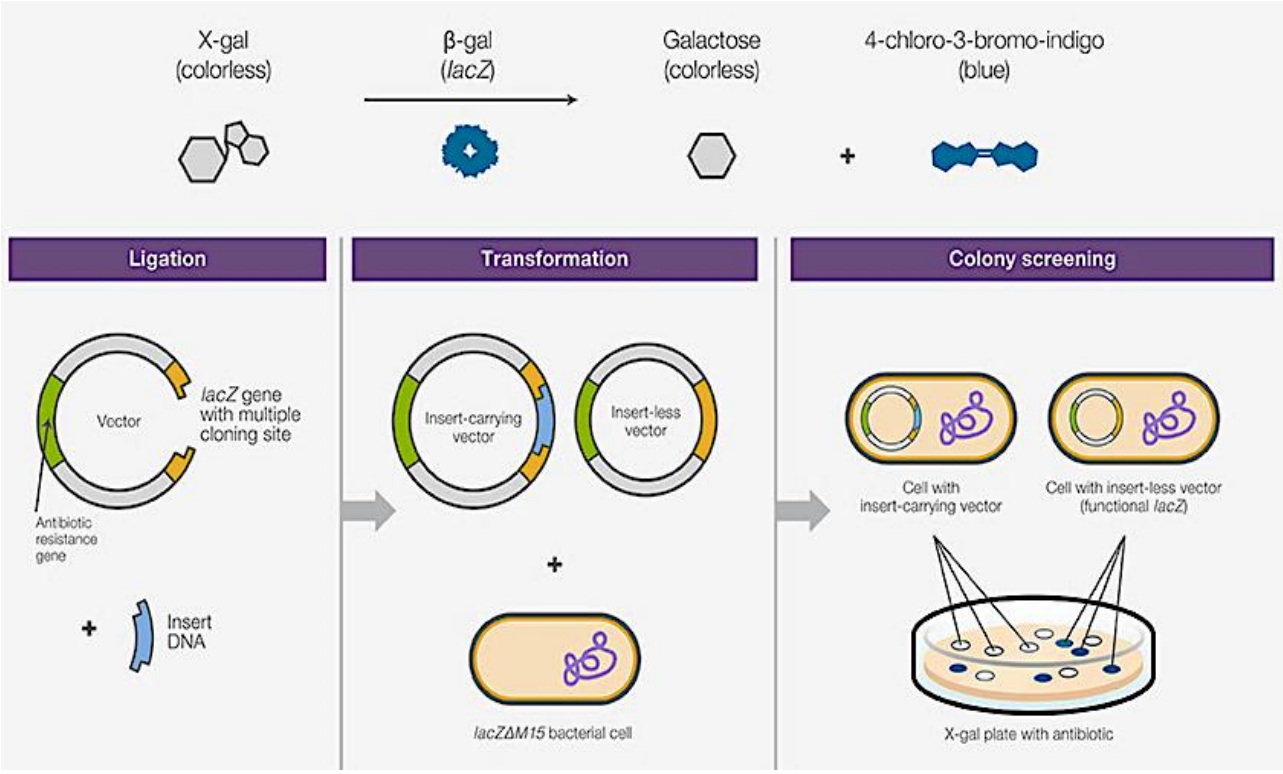


Figure 2.1 An overview of the TA-cloning procedure and blue/white screening of colonies by way of β-galactosidase activity. Figure taken from Thermo-Fisher Scientific: https://www.thermofisher.com/uk/en/home/life-science/cloning.html

2.3 Targeted deep sequencing and bioinformatics data analysis

A targeted myeloid panel of 33 genes frequently mutated in MDS/AML (**Table 2.2**) was employed to determine the secondary mutational repertoire across the sequential samples from the symptomatic patient (IV.10) compared with asymptomatic (III.7), WT (IV.9) and deceased MDS/AML (IV.6) family members (results in **Chapter 3.3** and **Appendix 2. Table 1**).

Table 2.2 List of 33 genes included in the targeted myeloid gene panel together with the regions sequenced.

Gene	Exons Targeted
ASXL1	12
BCOR	all
CALR	9
CBL	7+8+9
CEBPA	all
CSF3R	14-17
DNMT3A	all
ETV6	all
EZH2	all
FLT3	14+15+20
GATA2	all
GNAS	8+9
IDH1	4
IDH2	4
IKZF1	all
JAK2	12+14
KIT	2, 8-11, 13 +17
KRAS	2+3
MPL	10
NPM1	12
NRAS	2+3
PDGFRA	12,14,18
PHF6	all
PTPN11	3+13
RUNX1	all
SETBP1	4
SF3B1	12-16
SRSF2	1
TET2	all
TP53	all
U2AF1	2+6
WT1	7+9
ZRSR2	all

Bioinformatics analysis was performed by Dr Steven Best at King's College London. Target enrichment was achieved using an in-house TruSeq® Custom Amplicon (TSCA) design (Illumina, San Diego, USA). The targeted region consisted of a total of 71Kb represented by 295 amplicons. Pooled library targets were sequenced in batches of 24 samples on the MiSeq sequencing platform, using version 3.0 MiSeq sequencing reagents (Illumina, San Diego, USA). Minimum read depth threshold was 150 reads; lower limit of sensitivity was 5-10% variant allele frequency (VAF). All variants of unknown significance were eliminated (Tobiasson et al., 2016).

2.4 Gene expression analysis

Quantitative real-time PCR (RT-qPCR) analysis was performed to investigate global *GATA2* mRNA transcript levels across *GATA2*-mutated family members (results in **Chapter 3.4**). This generally involves: **1.** conversion of RNA to a more stable complementary DNA (cDNA) template via reverse transcription; **2.** PCR amplification of the cDNA and **3.** quantification of PCR products in real time using SYBR Green fluorescence.

2.4.1 RNA extraction and quantification

Total RNA was extracted from patient BM aspirates using the RNeasy Kit (Qiagen), as per the manufacturer's instructions. Approximately five million cells were collected from patient BM vials by centrifugation at 300g for 5 minutes. Cell pellets were then lysed in 350µL of Buffer (RLT) accompanied with 10µL/mL β-mercaptoethanol to denature RNAases. Cell lysates were then centrifuged at full-speed for 2 minutes using QIAshredder spin columns. 350µL of 70% ethanol was subsequently added to lysates to enable effective binding of RNA to the silica membrane. Samples were centrifuged at 8000g for 15 seconds using RNeasy Mini spin columns and were then washed with 350µL buffer RW1 at 8000g for 15 seconds.

An on-column DNase digestion step (80µL DNase I incubation mix containing 10µL of DNase and 70µL of RDD buffer) was performed and further incubated for 15 minutes at room temperature, to remove any remaining genomic DNA. After the DNase digestion step, the columns were washed with 350µL of buffer RW1 and with 500µL of buffer RPE followed by a final centrifugation step at full-speed for 2 minutes to remove any residual ethanol. Finally, the purified RNA was eluted intp 30µL of RNase-free water and stored at -80^{o}C. RNA concentrations were next determined using Qubit® 3.0 Fluorometer and the Quant-iT™ RNA Assay Broad Range Kit (Thermo Fisher Scientific, Waltham, USA) which is considered more sensitive in detecting dsDNA than Nanodrop.

2.4.2 cDNA synthesis

DNase-treated RNA was then reverse transcribed to single-stranded cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). In brief, 20µl of the reverse transcription (RT) reaction was prepared as described in **Table 2.3**. 10µl of the RT master mix was added to 10µl of RNA (~500ng) and the reverse transcription was carried out using the following thermal cycling conditions: (25°C for 10 minutes, 37°C for 2 hours and 85°C for 5 minutes).

cDNA samples were diluted in 480µL of RNase-free water (to reach a concentration of approximately 20ng/µl) which was then used as templates for the RT-qPCR reaction.

Table 2.3 Reagents used for cDNA synthesis.

Component	Volume (µl)
10X RT Buffer	2
10X RT random primers	2
MultiScribe Reverse Transcriptase	1
25X dNTP mix (100 mM)	0.8
Rnase free water	4
Rnase inhibitor	0.2
Total RNA (~500ng)	10
Final volume	**20**

2.4.3 Quantitative real-time PCR (RT-qPCR) using SYBR Green assay

Quantitative real-time PCR was performed on the synthesised cDNA using the SsoAdvanced Universal SYBR® Green assay (Bio-Rad) to measure *GATA2* expression levels (exonic RT-qPCR primers are listed in **Table 2.1**) relative to the expression of the housekeeping gene (*GAPDH*). Briefly, RT-qPCR reactions were run in triplicates and in a final volume of 10µl per reaction, consisting of: (4.6µl of cDNA template (20ng, prepared as described in **2.4.2**), 5µl of the SYBR Green PCR master mix and 0.2µl of forward and reverse *GATA2* exon-exon boundary primers (10µM concentration each). The universal SYBR Green comprises of all ingredients required for RT-qPCR reaction: (*Taq* Polymerase, dNTP mix with dUTP rather than dTTP, SYBR Green I DNA-binding dye and $MgCl_2$) to minimise pipetting errors and contamination risk.

These amplification reactions were performed in sealed 384-well plates on the CFX384 Touch™ Real-Time PCR Detection System (Bio-Rad). The cycling conditions include an initial polymerase activation and cDNA denaturation step at 95°C for 10 minutes followed by 40 cycles of cDNA denaturation to a single-stranded template at 95°C for 15 seconds, primer annealing at 60°C for 1 minute and cDNA final extension at 72°C for 30 seconds. Finally, a melting curve analysis was performed to check for the specificity of these amplifications.

2.4.4 RT-qPCR relative quantification and data analysis

Quantification and data analysis in RT-qPCR relies on the basic idea that the more copies of a particular gene are present in a particular sample, the fewer the amplification cycles needed to generate a fluorescent signal above threshold. The C_t (threshold cycle) therefore signifies the number of cycles required for a fluorescent signal to become detectable above a background sample and was used here to calculate the relative quantification. In order to control for differences in amplification efficiencies between samples, however, the expression levels of

target genes were normalised to the expression levels of a housekeeping gene (e.g. *GAPDH*) whose expression is considered constant regardless of cell type and experimental condition (Barber et al., 2005).

The $\Delta\Delta C_t$ method was employed to determine *GATA2* expression levels relative to *GAPDH* and using a healthy BM as a positive control sample (Livak and Schmittgen, 2001). First, the ΔC_t between *GATA2* and *GAPDH* was calculated for each sample ($\Delta C_t = Ct_{GATA2} - Ct_{GAPDH}$) and averaged across the three replicates. The average ΔC_t of each sample was then normalised to the ΔC_t of a positive control sample to give $\Delta\Delta C_t$ as follows: ($\Delta\Delta C_t = \Delta Ct_{sample} - \Delta Ct_{positive\ control}$). Finally, the normalised expression ratio ($2^{-\Delta\Delta Ct}$) gives the fold increase or decrease of *GATA2* expression between the samples and relative to the positive control sample.

2.5 RNA-sequencing and bioinformatics data analysis

High-throughput RNA-sequencing was performed to compare transcript profiles between *GATA2* monoallelic (IV.10-yr.1 and yr.3) vs. biallelic (IV.10-yr.4 and III.7) groups (results in **Chapter 3.5**). Total RNA was extracted using the RNeasy Mini Kit (Qiagen) and quantified using Qubit® 3.0 as previously mentioned and the integrity measured using the RNA 6000 nano assay kit on a 2100 BioAnalyzer (Agilent technologies, California, USA), where RNA integrity number (RIN) >7.0 was considered good quality RNA suitable for sequencing (**Table 2.4** and **Figure 2.2**). cDNA libraries were prepared using Poly-A selection and the TruSeq stranded mRNA library preparation kit (Illumina) according to the manufacturers' instructions and then sequenced to a minimum depth of 30 million (2x100bp) paired-end reads per sample (300 cycles of sequencing) on the Illumina HiSeq 4000, carried out at Oxford Genomics (Wellcome Trust Centre for Human Genetics, Oxford, UK).

Table 2.4 BioAnalyzer RNA quality control sample measurement. Table contains RNA concentrations by Qubit and RIN numbers by BioAnalyzer for each sample analysed.

Samples	BioAnalyzer RIN	Qubit 3.0 RNA concentration (ng/ul)
IV.10_yr.1	7.70	234
IV.10_yr.3	8.90	484
IV.10_yr.4	8.50	314
III.7_Asymptomatic	6.80	113
Healthy BM Control	7.50	280

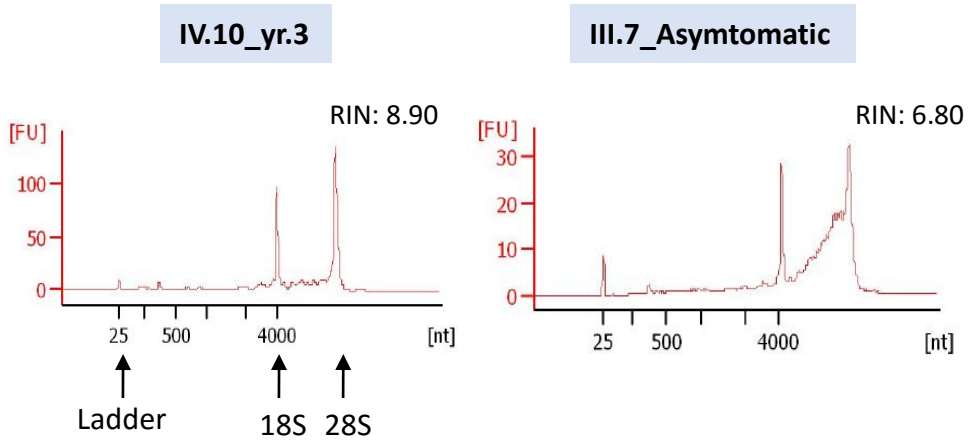


Figure 2.2 Example electropherograms produced using the Agilent 2100 BioAnalyzer. Good quality RNA displays a low threshold baseline (at the ladder) and a 28S rRNA peak twice as high as the 18S rRNA peak. Whereas a higher threshold baseline and a smaller 28S:18S ratio might indicate RNA degradation.

Bioinformatics data analysis was then completed by Dr Jun (Alex) Wang at the Barts Cancer Institute. Briefly, after FASTQ data quality check, raw reads were aligned to the reference genome (GRCh37) using STAMPY (Lunter and Goodson, 2011).The number of reads uniquely aligned (quality score q > 10) to the exonic region of each gene was counted using HTSeq based on the Ensembl annotation (Anders et al., 2015). Only genes that achieved at least one count per million

(CPM) reads in at least 2 samples (the number of samples analysed in each group) were included, leading to 17,555 filtered genes in total. Read counts were further normalised using the conditional quantile normalisation (CQN) method, accounting for gene length and GC content (Hansen et al., 2012). Differential expression (DE) analyses between *GATA2* biallelic vs. monoallelic groups were performed using the edgeR R package, employing the generalised linear model (GLM) approach (Robinson et al., 2010). The significantly DE genes were selected based on the false discovery rate (FDR) q-value of < 0.05 and the absolute $\log_2$ fold change value of > 1 (**Appendix 3. Table 2**). Gene set enrichment analysis (GSEA) was performed using the GSEA tool to identify canonical pathways, motifs and signature gene sets acquired from the Molecular Signatures Database (MSigDB-C2 v.5.0) (Subramanian et al., 2005). Top significantly enriched gene sets were selected based on FDR q-value of <0.05. The normalised enrichment score (NES) from each gene set was also shown. RNA-seq data have been deposited into Gene Expression Omnibus (GEO) under the accession number GSE104570.

2.6 Plasmid construct cloning, transfection and luciferase reporter assay

Luciferase reporter assay was employed to measure allele-specific differences in *GATA2* promoter transcriptional activity in our *GATA2* mutation carriers. First, luciferase reporter constructs were generated following subcloning of *GATA2* promoter 1 and 2 SNP alleles (4 plasmid clones) from the symptomatic patient (IV.10) and separately into the pGL2-Basic Vector (Promega) upstream of the firefly Luciferase gene (*Luc*) (**Figure 2.3-A**). The primers containing restriction enzyme (Xhol) used to amplify these promoter regions are listed in **Table 2.5**. DNA Plasmids were purified using the QIAprep Miniprep Kit (Qiagen) and insert identities and sequence orientation were verified by Sanger sequencing prior to transfection (**Figure 2.3-B**).

Table 2.5 ***GATA2*** **promoter 1 and 2 PCR primers for luciferase assay** – contacting restriction site XhoI.

Primer	Forward	Reverse
Promoter 1_Luc_XhoI	CAC**CTCGAG**GTGGCTCCGAGAAATGGGAA	TTC**CTCGAG**CGCTGACTGGTTGAATCCCT
Promoter 2_Luc_XhoI	CAC**CTCGAG**CACACTTGGCGCCAGATACA	TTC**CTCGAG**CCGATCTCCGGGCTAGAAGT
Vector GL primer	TGTATCTTATGGTACTGTAACTG	CTTTATGTTTTTGGCGTCTTCCA

1µg of DNA firefly luciferase vector containing either one of the four alleles from *GATA2* promoter 1 and 2 SNPs were transiently co-transfected into HeLa cells (1.25 x 10^5 per well in a 6-well plate) together with 0.5µg of DNA Renilla luciferase control vector pRL-CMV (Promega) using the Lipofectamine™ 2000 transfection protocol (Invitrogen). 1µg of empty vector was transfected as well to act as a control. Cells were lysed 48 hours after transfection in accordance with the Dual-Luciferase® Reporter Assay System (Promega) as previously described (Sherf et al., 1996).

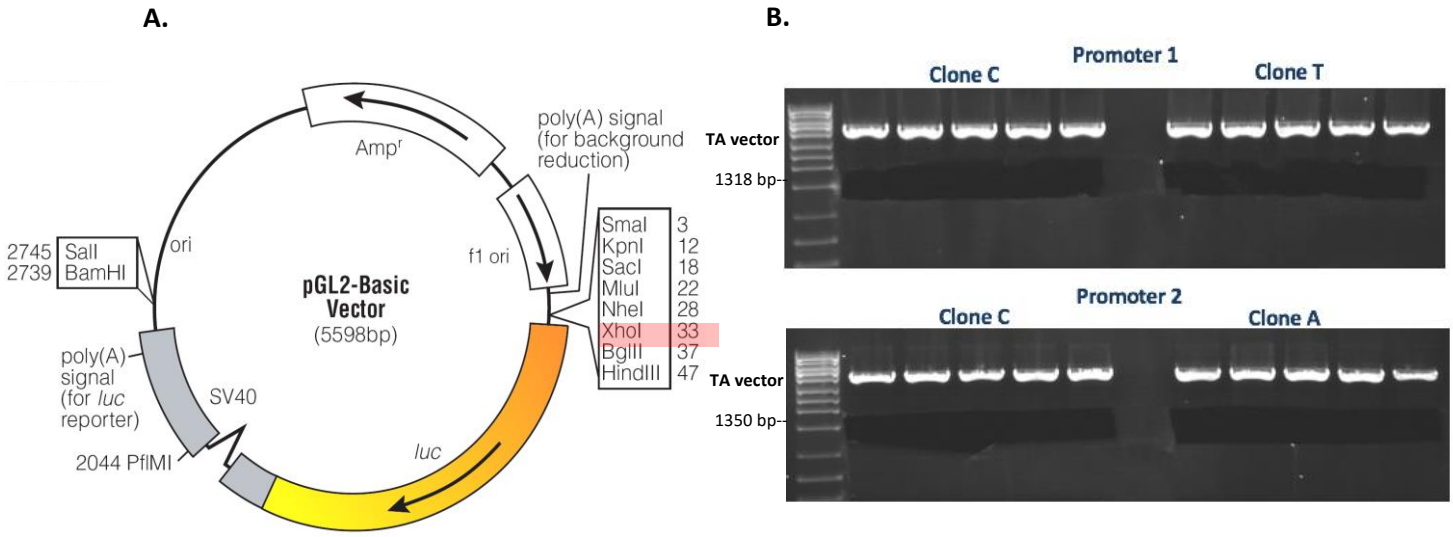


Figure 2.3 Luciferase reporter assay. A. Map showing the promotorless firefly luciferase vector (pGL2-Basic - Promega) used in this study and the region where *GATA2* promoter fragments were inserted into XhoI restriction site (highlighted in red) in the multiple cloning site upstream of the *Luc* gene. **B.** PCR gels showing separated TA vector (upper lane) and DNA inserts (promoter 1 and promoter 2) (lower lane) on 0.8% agarose gel after digestion with XhoI. Bands corresponding to promoter 1 clones [C/T] (1318bp) and promoter 2 clones [C/A] (1350bp) were excised for DNA purification prior to cloning into the luciferase vector.

The activities of Firefly and Renilla luciferases were measured sequentially from a single sample; the Firefly luciferase reporter was measured first by adding Luciferase Assay Reagent II (LAR II) to generate a stabilised luminescent signal. The reaction was then quenched subsequently after addition of Stop & Glo® reagent to activate the Renilla luciferase reaction (**Figure 2.4**). This is achieved using the luminescence function on the FluoStar Optima plate reader (BMG Labtech, Germany).

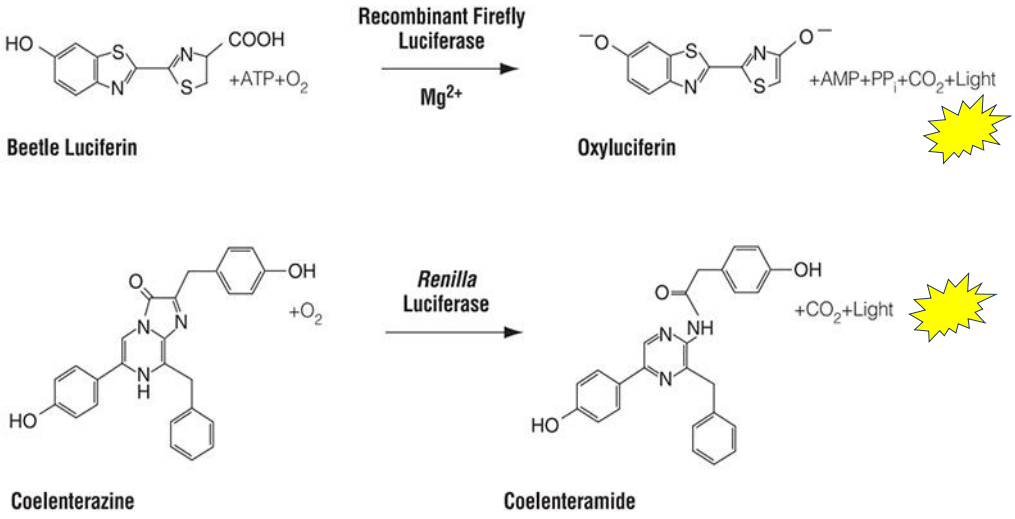


Figure 2.4 Bioluminescent reactions catalysed by Firefly (*Photinus pyralis*) and Renilla (*Renilla reniformis*) luciferases. For Firefly luciferase, light emission is achieved by oxidation of Beetle Luciferin in a reaction that requires ATP, Mg^{2+} and O_2 whereas the luminescent reaction catalysed by Renilla luciferase utilises O_2 and Coelenterazine. Figure reproduced from the Promega dual-luciferase reporter assay system protocol and technical manual.

The luciferase reporter assays were performed in triplicates (5 independent experiments) and the final results were expressed as a percentage relative luciferase activity calculated by normalising the ratio of Firefly luciferase to Renilla luciferase luminescence and a further normalisation to an empty vector control.

2.7 DNA methylation analysis

2.7.1 Bisulphite DNA modification

Genomic DNA of *GATA2*-mutated family members was bisulphite-converted to quantifiably measure DNA methylation at a single CpG level. Essentially, bisulphite conversion involves the deamination of cytosine residues by sodium bisulphite, thereby converting non-methylated cytosines into uracil (which is then PCR amplified into thymines) whilst keeping the methylated cytosines intact and thus protected from deamination (Frommer et al., 1992) (**Figure 2.5**). The ratio of methylated cytosines compared to the total number of cytosines and/or thymines in a given converted sequence can therefore give an indication of the level of DNA methylation at that particular locus.

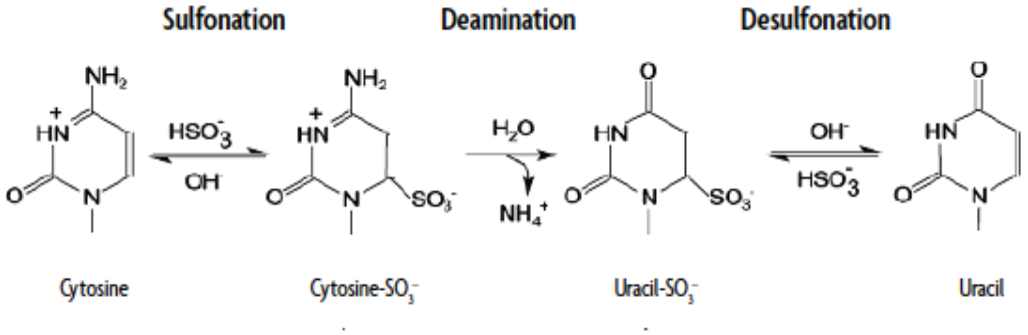


Figure 2.5 Schematic of bisulphite DNA conversion reaction of non-methylated cytosine into uracil. The first two steps (sulfonation and deamination) are performed at low pH while the final step (desulfonation) is performed under high pH conditions.

Using the Bisulphite Conversion Kit (Active Motif), genomic DNA (~500ng) was first treated with Proteinase K (20mg/ml), to improve the efficiency of the conversion reaction, and incubated at 50°C for 30 minutes. Following incubation, samples were then rapidly heat-denatured upon addition of the bisulphite conversion buffer (120µl) and hydroquinone (7µl) (protects DNA from strand breakage caused by depurination) using the following cycling conditions: 95°C for 30

seconds, 58°C for 20 minutes and 3 cycles of (95°C for 10 seconds and 58°C for 20 minutes). Following this 1.5-hour DNA conversion reaction, DNA samples together with 500µl of DNA binding buffer were added onto DNA purification columns and undergone several washes and centrifugation using ethanol-containing wash buffers. Next, 200µl of desulfonation buffer was added and the mixture was incubated for 20 minutes at room temperature. DNA is then washed and eluted from the columns by addition of 30µl of elution buffer. To check for the efficiency of the conversion, DNA samples were amplified using positive control conversion-specific PCR primer set (Active Motif) and run on a 1.5% agarose gel. Samples were considered to be successfully bisulfite-converted if they showed a PCR product.

2.7.2 Methylation-specific PCR (MSP)

Bisulphite-treated DNA was then PCR amplified using specific methylated (M) or unmethylated (UN) primer pairs covering *GATA2* promoter 1 and 2 regions (**Table 2.1**) and the PCR products were visualised using 1.5% agarose gel electrophoreses (as detailed in section **2.2.3**) to assess for global DNA methylation patterns between symptomatic and asymptomatic carriers.

2.7.3 Bisulphite-specific PCR (BSP), cloning and sequencing

Bisulphite-treated DNA was PCR amplified using primers encompassing a 200-bp region of *GATA2* second promoter and overlapping a CpG island (containing ~20 CpGs) (**Table 2.1**). The PCR products were then cloned into pCR2.1® vectors using the Original TA-cloning Kit (Invitrogen) and white colonies (approximately 25-30 clones per sample) were selected and DNA purified as mentioned previously in section **2.2.6** and sequenced using the vector M13 primers (**Table 2.1**) to measure allelic-specific differences in *GATA2* promoter methylation between mutant and WT alleles across the symptomatic patient's (IV.10) different time-points.

2.8 Chromatin immunoprecipitation (ChIP) assay

ChIP was performed using the ChIP-IT High Sensitivity Kit (Active Motif) and following the manufacturer's instructions, to assess allele-specific differences in the deposition of chromatin marks (H3K4me3 and H3K27me3) in our patient samples (**Figure 2.6**). Briefly, the symptomatic patient (IV.10) fresh BM samples (~7 million cells) across three different time-points (y.1, yr.3 and yr.4) were each subject to 5 minutes of formaldehyde fixation to cross-link and maintain protein/DNA interactions. Fixed chromatin was then sonicated for ~15 cycles to an average length of 200bp using the UCD-300 Bioruptor plus sonication system (Diagenode, Belgium). 2µg of H3K4me3 or 1µg of H3K27me3 antibodies (Diagenode) were added separately to sheared chromatin and reactions were incubated on a rotator overnight at 4°C. Chromatin bound to either of these antibodies was then immunoprecipitated using Protein G Agarose beads (Active Motif). ChIP reactions were next subject to several washes by gravity filtration after which proteins were removed and ChIP-enriched DNA was purified (in a process called reversal cross-linking) using proteinase K treatment and DNA purification columns. H3K4me3 or H3K27me3 ChIP-enriched DNA fragments were analysed by direct PCR and Sanger sequencing within *GATA2* promoter regions using primer pairs with smaller amplicon sizes (See **Table 2.1** for primer sequences).

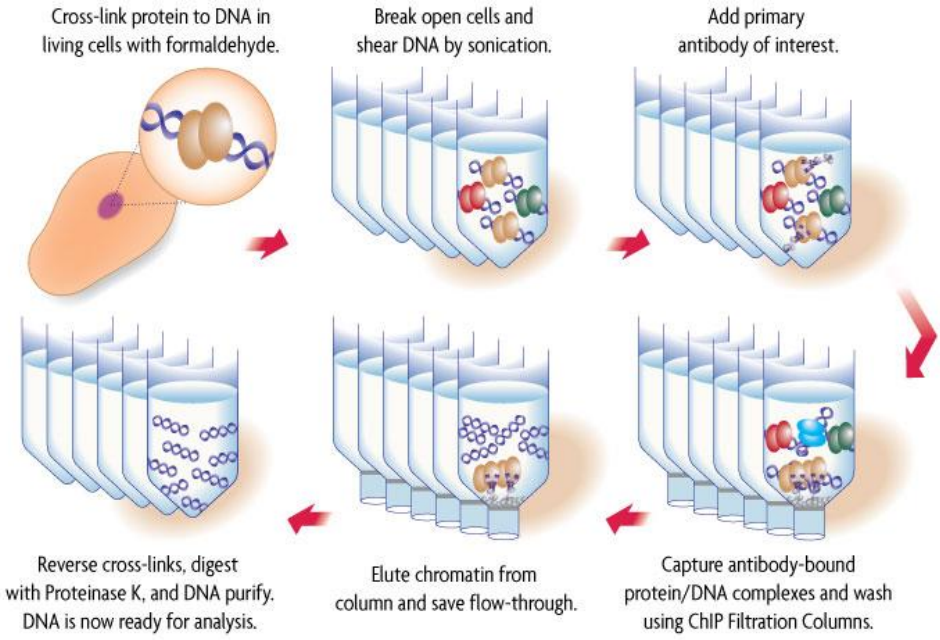


Figure 2.6 Schematic workflow of chromatin immunoprecipitation (ChIP). Figure obtained from Active Motif instruction manual (https://www.activemotif.com).

2.9 Statistical analysis

Statistical significance for relative levels of gene expression in RT-qPCR, luciferase reporter experiments, proportions of methylated CpGs in bisulphite cloning experiments and sequencing peak heights in ChIP experiments was determined at $p \leq 0.05$ (*), $p \leq 0.01$(**) and $p \leq 0.001$(***), calculated using *t*-test with Bonferroni correction. Data are expressed as mean ± standard error of the mean (SEM) of 3 to 5 independent experiments performed in triplicates and plotted using the GraphPad Prism version 8.0 software (GraphPad Software Inc., La Jolla, CA, USA).

Chapter 3. Results

Investigating the Molecular Mechanisms Underlying Reduced Penetrance of Germline *GATA2* Mutations

"Illness is the night-side of life, a more onerous citizenship. Everyone who is born holds dual citizenship, in the kingdom of the well and in the kingdom of the sick. Although we all prefer to use only the good passport, sooner or later each of us is obliged, at least for a spell, to identify ourselves as citizens of that other place."

– Susan Sontag

3.1 Background and Rationale

As introduced in **Chapter 1**, GATA2 is a master regulator of haematopoiesis, expressed in early BM progenitors and plays a key role in their differentiation and survival (Collin et al., 2015). While *GATA2* is mutated at low frequencies (~4-7%) in sporadic leukaemia, germline mutations in *GATA2* are commonly identified in autosomal dominant familial MDS/AML, particularly in *de novo* paediatric MDS (7-15%) (Wlodarski et al., 2016) and predispose its carriers to protean and overlapping clinical manifestations (including Emberger, MonoMAC and DCML deficiency) highlighting the pleiotropic nature of this transcription factor.

Following the initial report of germline *GATA2* mutations in four MDS/AML families by Hahn and colleagues in 2011, more than 30 pedigrees have been reported worldwide so far (**Table 1.5**). We have noticed that in a proportion of these families, especially those carrying germline missense mutations, there is a striking evidence of reduced penetrance (50-70%) where multi-generation mutation carriers are split between symptomatic and asymptomatic family members. The molecular basis of such occurrence however has not been thoroughly explored. We hypothesised that these germline *GATA2* mutations are unable to initiate disease solo but rather act synergistically with subsequent molecular events in order to facilitate clonal proliferation and expansion.

- The findings reported in this chapter and the following one (**3** and **4**) were published in and figures reproduced from (Al Seraihi et al., 2018) in *Leukemia* – Springer Nature (manuscript provided in **Appendix 1.**).

3.2 Clinical profile of a *GATA2*-mutated MDS/AML pedigree with reduced penetrance

Analysis of MDS/AML families harboring (p.Thr354Met (p.T354M), c.1061C>T) *GATA2* mutations displayed significant intra- and inter-familial variations in disease latency, clinical phenotype, and penetrance. These observations suggest that individuals require additional co-operating aberrations for the development of overt malignancy within the context of a shared germline mutation.

To investigate this hypothesis further, we examined a five-generation MDS/AML family who presented to St. Bartholomew's hospital in London with varying symptom manifestations and ages of disease onset and where germline heterozygous p.T354M *GATA2* mutations were detected in six individuals spanning two generations (III.1, III.5, III.7, IV.1, IV.6 and IV.10) (Bodor et al., 2012)(**Figure 3.1-A**).

As illustrated in the clinical timeline (**Figure 3.1-B**), the family came into our attention when two first-degree cousins (IV.1 and IV.6) presented within one week from each other with high-grade MDS transforming to AML aged 18 and 23 years, respectively. IV.1 presented with multiple plantar warts and 7% myeloblasts in his diagnostic BM consistent with refractory anaemia with excess blasts-1 (RAEB-1). He had an absolute monocytopenia [0.0 x 10^9/L] at the time of presentation and underwent a 1-antigen mismatched unrelated donor HSCT but unfortunately passed away two years later due to transplant-related complications and relapsed disease.

The second cousin's (IV.6) diagnostic BM displayed trilineage dysplasia with 17% myeloblasts (RAEB-2) albeit his monocyte count was within normal range [0.3 x 10^9/L]. He underwent intensive chemotherapy and obtained complete remission but with persistent dysplastic features in the BM. A second cycle of chemotherapy was further complicated by long-lasting anaemia and thrombocytopenia; he relapsed seven months following presentation and received re-induction chemotherapy and a haplo-identical HSCT from his mother (III.6) but developed GvHD, severe infections and died of pneumonic sepsis four months following allogeneic HSCT. Cytogenetic analysis confirmed monosomy 7 in both cousins and the *GATA2* mutation was identified in the diagnostic tumour DNA samples from IV.1 and IV.6 as well as in a remission DNA sample from IV.6, supporting a germline origin (Bodor et al., 2012). The first part of this project therefore involved validating these mutations by Sanger sequencing, as part of the initial genetic profiling performed in this family (**Figure 3.1-C**).

Ten years after the cousins' (IV.1 and IV.6) presentation, their first cousin (IV.10) presented at 31 years old with recurrent minor infections and severe leukopenia (monocytopenia [0.1x10^9/L] and neutropenia [0.8x10^9/L]) with mild macrocytosis and normal karyotype, haemoglobin and platelet counts. These clinical parameters subsequently stabilised (normal monocyte count, neutrophils [>1x10^9/L]) three years (yrs.) after presentation (**Figure 3.2**). It is worth mentioning here that IV.10 has not undergone any intensive chemotherapy regimen during this time period and continues to remain under close surveillance where her disease progression and blood counts are routinely monitored. Fortunately, all four of her children (V.1-4) have inherited her wildtype (WT) *GATA2* allele. Similarly, healthy family members from the fourth generation (IV.7, IV.8, and IV.9) were screened for the mutation and all have a WT *GATA2* configuration.

The paternal grandmother (II.2) of IV.10 as well as her paternal great-uncle (II.3) and great-grandmother (I.2) all were reported to have succumbed to leukaemia (ages of disease onset were 53, 24, and 53 years old, respectively) although it is not known whether they were too carriers of the *GATA2* mutation (genomic DNA material was not available).

However quite remarkably, not only did the presence of germline *GATA2* mutations correlate with early age of disease onset in the fourth generation (IV.1_18yr., IV.6_23yr. and IV.10_31yr.), but the parental third generation *GATA2* mutation carriers (III.1, III.5, and III.7) remain haematologically normal and symptom-free, with no history of recurrent infections despite being in their mid-late 60s. This occurrence therefore spurred us to investigate what is protecting those asymptomatic carriers and indeed what is driving the onset of overt disease in the symptomatic ones.

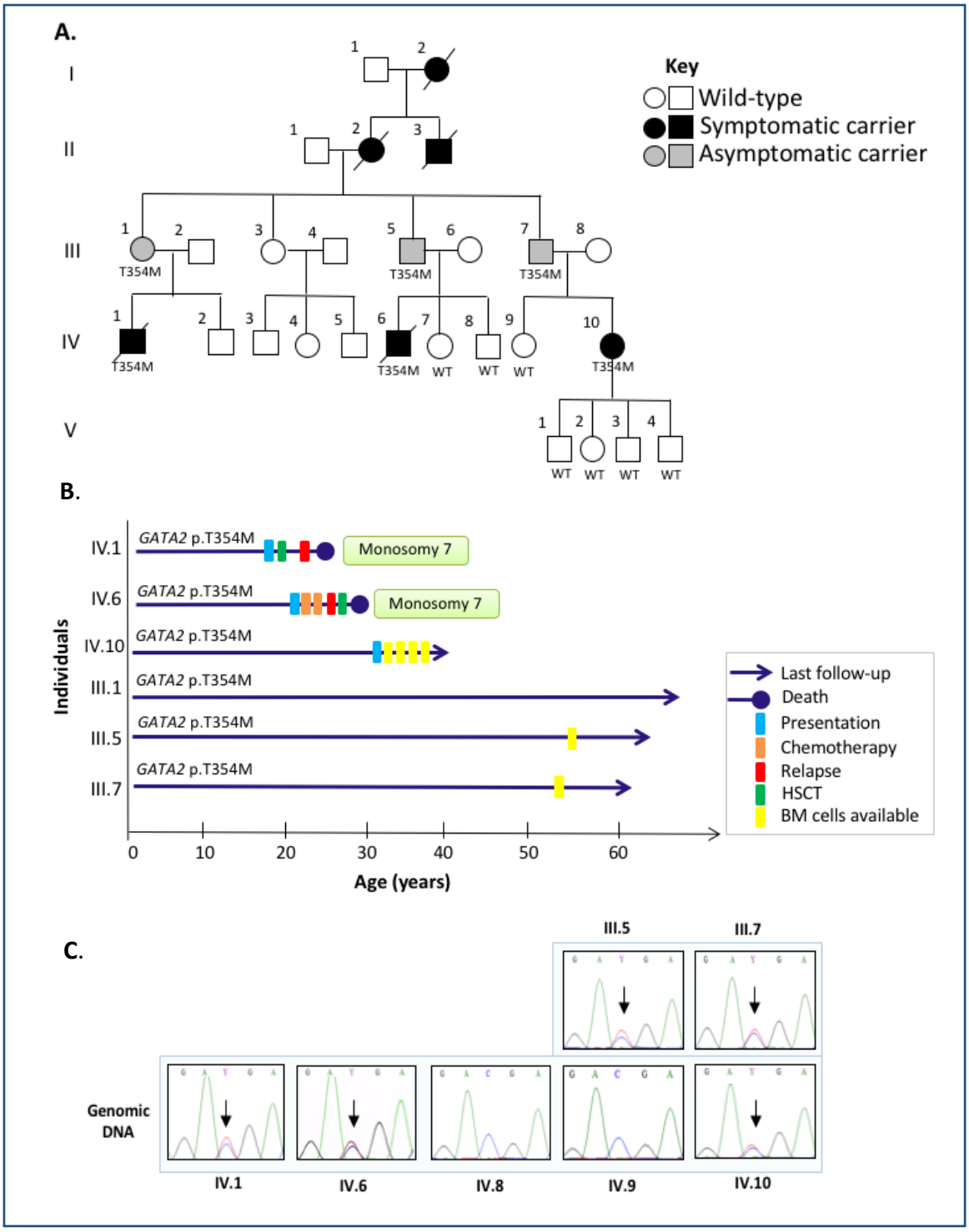


Figure 3.1 A. Genogram of the *GATA2*-mutated MDS/AML pedigree with reduced penetrance. Squares denote males and circles denote females. A strikethrough line indicates a deceased family member (pedigree redrawn from Bodor et al., 2012). **B.** Clinical timeline of disease events in *GATA2*-mutated family members showing deceased cousins (IV.1 and IV.6) harbouring secondary aberrations (monosomy 7) whilst the third cousin (IV.10) remains under routine clinical follow-up. Third generation mutation carriers (III.1, III.5 and III.7) remain asymptomatic well into their 60s. Yellow squares represent BM cells available for this study. **C.** Sanger sequencing chromatograms demonstrating heterozygous germline *GATA2* mutations (c.1061C>T, p.T354M) validated in five family members from the third and fourth generations with variable disease latency and phenotype. No DNA was available from the symptomatic individuals (I.2, II.2 and II.3) nor from other family members who were not mentioned.

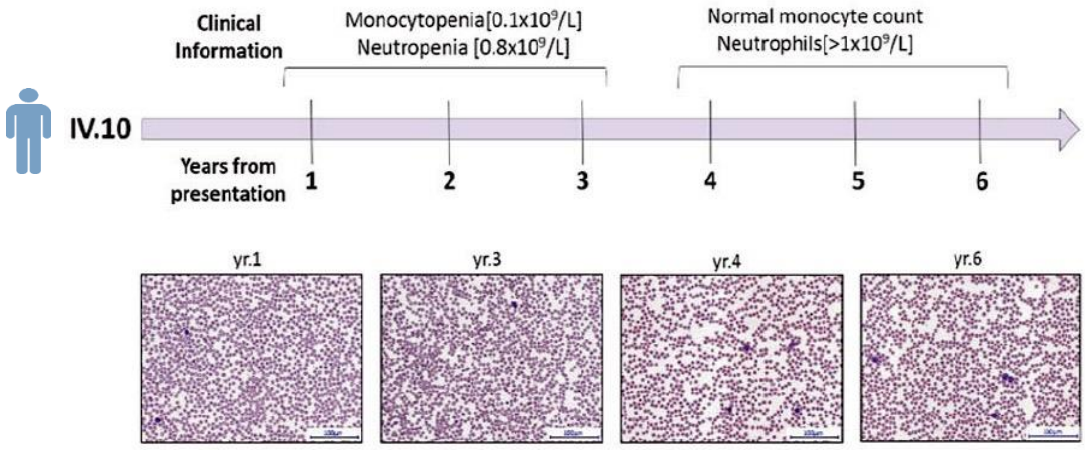


Figure 3.2 Clinical timeline of IV.10 symptomatic patient showing the changes in clinical parameters over the course of disease presentation. Photographs of PB smears from IV.10 (yr. 1, 3, 4 and 6) stained with May-Grünwald Giemsa staining. Magnification: x20.

3.3 Acquisition of somatic *ASXL1* mutations in symptomatic vs. asymptomatic carriers

Our initial assumption was that disease penetrance was governed by the acquisition of additional secondary mutations, akin to the group's previous work on familial *CEBPA*, where the onset of overt leukaemia accompanied the acquisition of secondary C-terminal mutations (Smith et al., 2004, Tawana et al., 2015). In order therefore to define the pattern of secondary genetic events across *GATA2*-mutation carriers that may account for the clinical heterogeneity and the reduced penetrance observed in this family, targeted deep sequencing using a myeloid panel of 33 genes frequently mutated in MDS/AML (e.g. *DNMT3A*, *RUNX1*, *TET2* and *WT1*) was performed in select members of the family with a minimum depth coverage of ~1500x (see **Chapter 2.3** – **Table 2.2** for a list of these 33 genes). Precisely, BM DNA samples from four family members were tested: one asymptomatic parent (III.7), one deceased MDS/AML cousin (IV.6), one healthy WT sibling (IV.9) and across three time-points (yr. 1, 4, and 6) from the symptomatic patient (IV.10), where sequential BM material became available through routine clinical follow-up (**Figure 3.2** and **3.3-A**).

As illustrated in **Figure 3.3-B**, the high VAFs of *GATA2* p.T354M (range 46–51%) across mutation carriers reflect the germline origin of the mutation. Notably, while no acquired somatic mutations were detected in asymptomatic family members, all affected cousins analysed acquired an identical frameshift *ASXL1* lesion (p.Gly646TrpfsTer12, c.1934dupG) (**Figure 3.3-B**). Indeed, *ASXL1* mutations have been linked as a secondary genetic event in germline *GATA2*-mutated pedigrees (West et al., 2014b, Wang et al., 2015), however, it was noteworthy that *ASXL1* mutation was not clonal in IV.10 but critically remained stable (VAF range 12–6%) over a 6-year monitoring period and was therefore not considered a primary driver for the development of overt MDS/AML.

Apart from the *ASXL1* mutation, no other somatic mutations were detected in the 33-myeloid genes assessed in the symptomatic cousins (except for few shared non-pathogenic SNPs (e.g. in *TET2* and *TP53*), listed in **Appendix 2. Table 1**). Moreover, and as expected, *ASXL1* mutation was not identified in the remission sample of the cousin with MDS/AML (IV.6), confirming that it represents an acquired secondary mutation.

While the molecular co-occurrence of *ASXL1* and *GATA2* mutations has been proposed as one mechanism for driving the onset and severity of disease symptoms (West et al., 2014b), the low VAF of *ASXL1* mutation and the steady improvement in haematopoiesis at IV.10 later follow-up (yrs. 4 and 6) (**Figure 3.2**) suggested that a combination of *GATA2* and *ASXL1* mutation is perhaps still not sufficient on its own to promote clonal expansion and leukaemic transformation. This indicates that other molecular events, possibly preceding the acquisition of *ASXL1* mutation, might trigger the patient's initial symptoms.

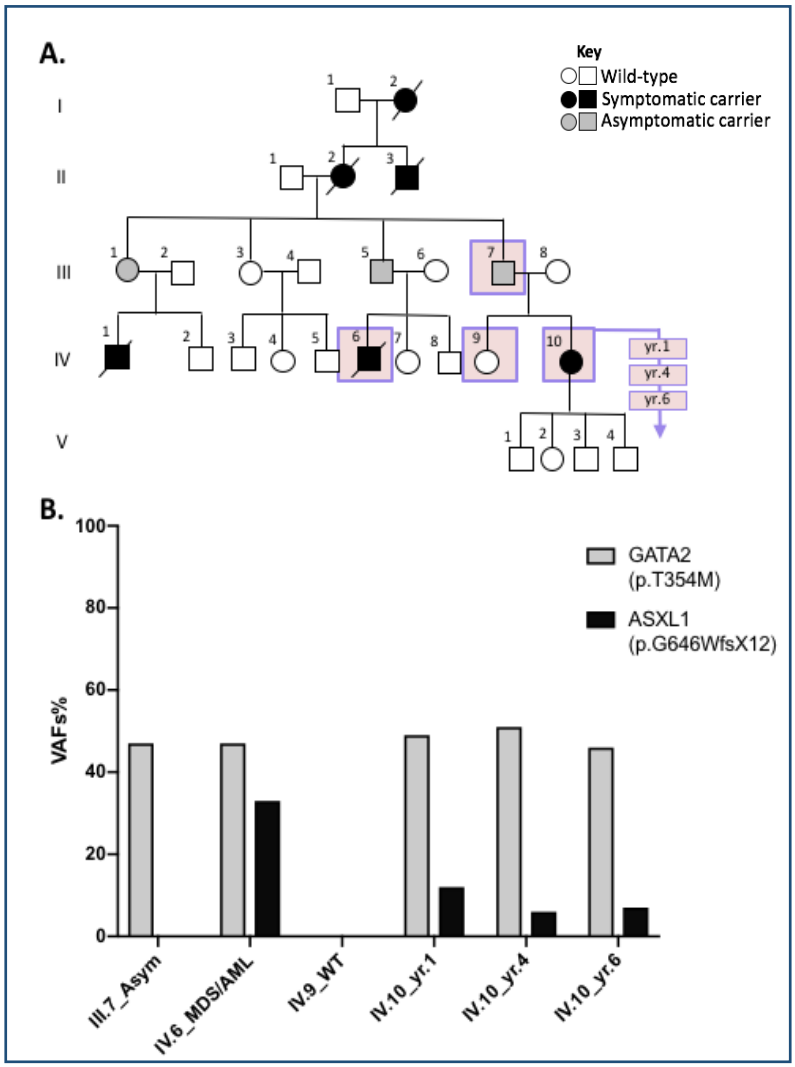


Figure 3.3 Identification of secondary *ASXL1* somatic mutations in *GATA2*-mutated familial MDS/AML. **A.** Genogram of the *GATA2*-mutated family highlighting samples included in the targeted deep sequencing myeloid gene panel. **B.** Bar chart illustrating germline *GATA2* (p.T354M) and acquired *ASXL1* (p.G646WfsX12) mutations identified in each sample together with their VAF levels. It is worth mentioning here that the MDS/AML cousin (IV.1) also acquired *ASXL1* mutations (verified by Sanger sequencing) but was not included in the myeloid gene panel.

3.4 Monoallelic *GATA2* expression differentiates between symptomatic and asymptomatic carriers and correlates with clinical parameters

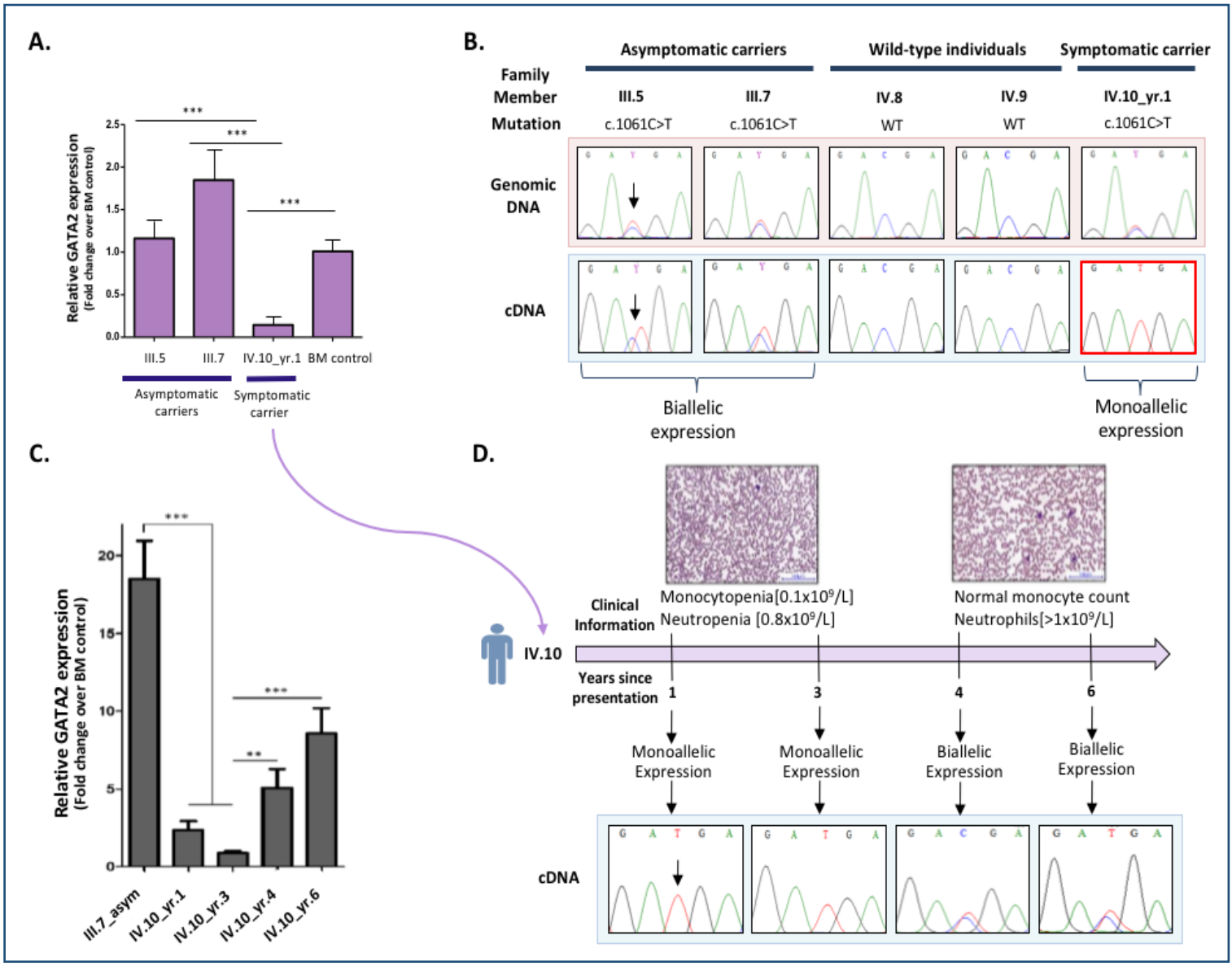
We next focused our attention on the *GATA2* locus and investigated whether disease symptoms are influenced by endogenous levels of *GATA2. GATA2* mRNA levels were first compared amongst family members using quantitative RT-PCR of BM material which showed total *GATA2* expression to be significantly lower in the symptomatic (IV.10_yr.1) compared with asymptomatic carriers (III.5 and III.7) (**Figure 3.4-A** and methods **Chapter 2.4**). We then wondered whether this global reduction in expression was accompanied by differences in the WT vs. mutant allelic expression; indeed, Sanger sequencing of *GATA2* BM cDNA template revealed striking allele-specific expression (ASE) favouring the mutant allele (**T**) with an absence (or silencing) of the WT allele (**C**) expression in the symptomatic patient (IV.10), which contrasted with biallelic expression in asymptomatic family members (III.5 and III.7) (**Figure 3.4-B**). The observation of differential allelic-specific expression, favouring the mutant allele and leading to silencing of the WT allele, provided a rationale for explaining the differences between symptomatic and asymptomatic carriers in this family and suggests that the protection effect seen in the asymptomatic family members might reflect on the maintenance of the WT allele expression.

As this suggested that changes in mutant:WT *GATA2* allelic expression ratio may account for the reduced penetrance observed in this family, we had the opportunity to sequentially monitor *GATA2* expression in IV.10 over a 6-year disease period where we collected BM material at 4 different intervals (yr.1, 3, 4 and 6), as part of the patient's routine clinical management. When we compared global *GATA2* expression over this 6-year monitoring period (depicted in **Figure 3.4-C**), we noticed that later time-points of IV.10 (yr.4 and 6) demonstrated increased global

GATA2 expression than earlier periods (yr.1 and 3), as assessed by RT-qPCR (described in methods **Chapter 2.4**). Remarkably, this upward shift coincided with reactivation of the WT allele (**C**) expression, demonstrated by the presence of a double cDNA sequencing peak (corresponding to both mutant (**T**) and WT (**C**) alleles) and a persistent improvement in haematological parameters (increase in neutrophil counts to >1x10^9/L), noted 3 years after presentation (yr. 4 and 6) and in the absence of any treatment intervention (**Figure 3.4-D**).

These monoallelic vs. biallelic *GATA2* expression findings (based on qualitative Sanger sequencing trace chromatograms) were further validated by cDNA cloning of a 175bp fragment of *GATA2* exon 5 region encompassing the p.T354M germline heterozygous mutation from IV.10 (symptomatic), III.7 (asymptomatic) and IV.9 (WT) BM samples into pCR®2.1 vectors and subsequent Sanger sequencing of individual clones (an average of 25 clones were analysed per sample) to quantitatively measure the proportion of mutant (**T**) to WT (**C**) alleles expressed in each sample analysed (see methods **Chapter 2.2.6**). In particular, this demonstrated an allelic *GATA2* mutant:WT clone ratio of 80.4**:**19.6% in IV.10 earlier time-points (yr. 1 and 3) compared with an allelic ratio of 50.2**:**49.8% in later periods (yr. 4 and 6) (**Figure 3.5-A** and **B**) confirming our cDNA Sanger sequencing results.

Altogether, even though RNA material was not available to test *GATA2* allelic expression at early time points of disease development in the two deceased cousins (IV.1 and IV.6), we believe that the silencing of the WT *GATA2* allele and ASE of the mutant allele, leading to an almost complete depletion of GATA2 function, might indeed be one of the initial steps required to induce symptom manifestations (e.g. cytopenia) at the very early stages of the symptomatic patient's disease development.



<u>Figure 3.4</u> Investigating *GATA2* mRNA expression. A. *GATA2* global expression measured by RT-qPCR of BM samples and normalised to healthy BM control. Bar chart demonstrating downregulation in IV.10_yr.1 *GATA2* expression compared with III.5 and III.7 asymptomatic carriers. No RNA material was available from the deceased MDS/AML cousins (IV.1 and IV.6) nor from the asymptomatic carrier (III.1). An average of 3 independent experiments is shown. Statistical significance was determined at *p ≤ 0.05, **p ≤ 0.01 and ***p ≤ 0.001 using a *t*-test with Bonferroni correction. Error bars represent standard error of the mean (SEM). **B.** Allele-specific expression of the *GATA2* mutant allele (T) in the symptomatic (IV.10_yr.1) vs. asymptomatic (III.5 and III.7) carriers, as measured by cDNA Sanger sequencing of BM samples. **C.** Bar chart demonstrating a downregulation in IV.10_yr.1-3 global *GATA2* expression compared with IV.10_yr.4-6 as measured by RT-qPCR. An average of 5 independent experiments is shown. Statistical significance was determined at *p ≤ 0.05, **p ≤ 0.01 and ***p ≤ 0.001 using a *t*-test with Bonferroni correction. Error bars represent SEM. **D.** Correlation of monoallelic *GATA2* expression with disease state across the 4 time-points studied in IV.10, with reactivation of the WT allele (C) expression noted 3 years after presentation (yr.4 and 6), concurrent with improvements in clinical parameters.

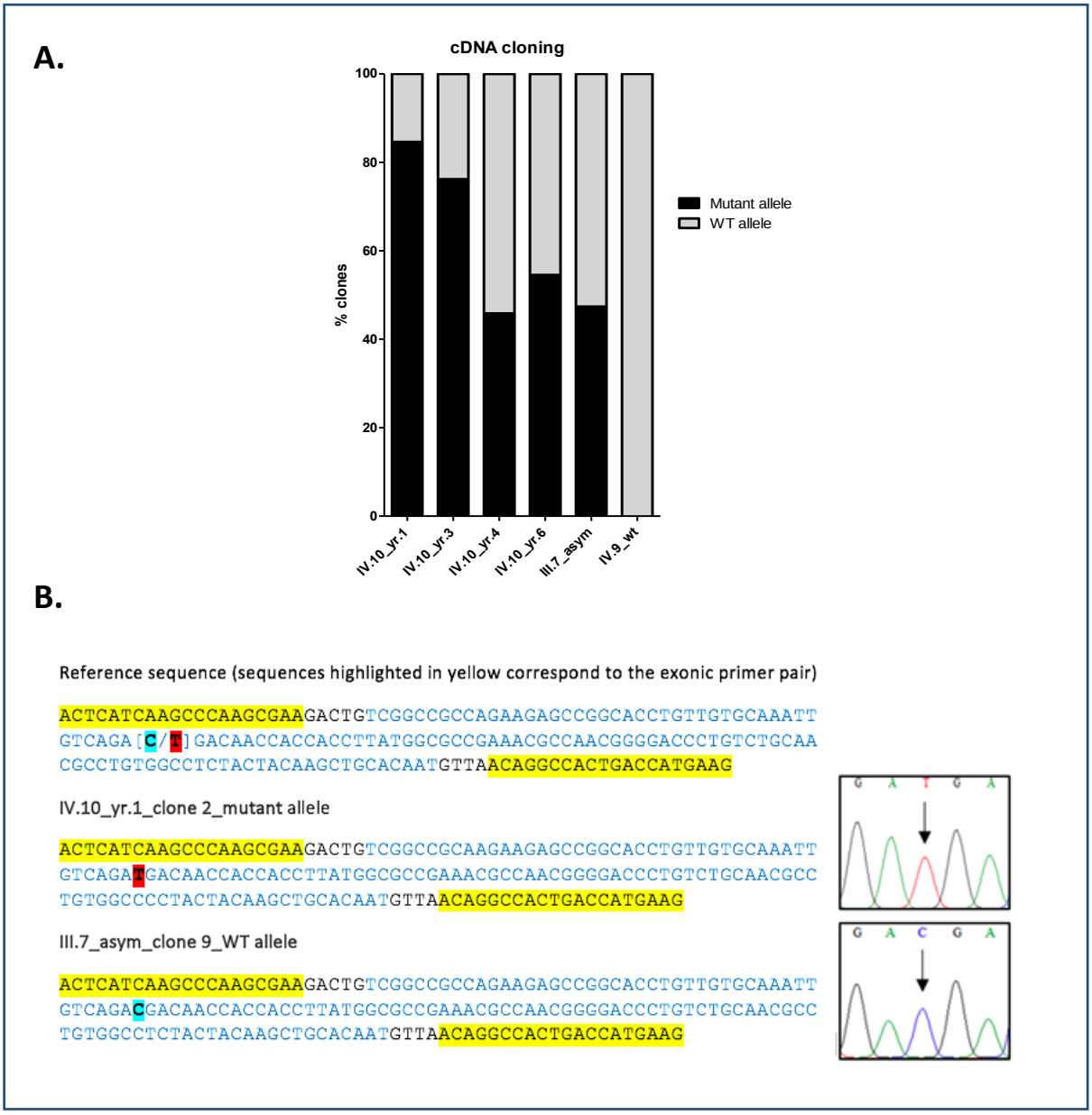


Figure 3.5 cDNA cloning to quantitatively measure *GATA2* mRNA allelic expression. A. Bar chart showing the percentage of mutant and WT allele clones across the 4 time-points of IV.10 (yr.1, 3, 4 and 6) and in III.7 (asymptomatic carrier) and IV.9 (WT sibling) based on cDNA cloning of *GATA2* and Sanger sequencing of individual clones (an average of 25 clones were analysed per sample).
B. A representative example of mutant (T) and WT (C) allele clones together with their corresponding Sanger trace chromatograms. Sequences in blue represent *GATA2* exon 5 where the germline mutation (c.1061C>T, p.T354M) resides. Sequences highlighted in yellow represent the primer pair used to amplify this region prior to cloning the cDNA PCR products into pCR®2.1 TA-vectors (see **Chapter 2.2.6** for a detailed method description).

3.5 *GATA2* monoallelic samples display unique transcript profiles compared to their biallelic counterparts

Based on the aforementioned results, we now have two entities: *GATA2* monoallelic and *GATA2* biallelic expressing samples. The next question we asked was how similar or different are those two biological entities and whether *GATA2* mutant ASE has an impact on the transcriptome driving the onset of disease symptoms. To this end, we performed RNA-seq analysis with a view of examining downstream biological features distinctive of each of these two groups. We tested five RNA samples derived from patients' BMs: two *GATA2* monoallelic (IV.10_yr.1 and 3), two *GATA2* biallelic (IV.10_yr.4 and III.7) and one healthy BM control with WT *GATA2*. Following an in-house sample quality control (see **Chapter 2.5**), cDNA libraries were prepared using Poly-A selection and then sequenced to a minimum depth of 30 million (2x100bp) paired-end reads per sample, carried out at Oxford Genomics. Bioinformatics analysis was then completed by Dr Jun (Alex) Wang at the Barts Cancer Institute.

While it is important to acknowledge that the number of samples available is few, as shown in **Figure 3.6-A**, an unsupervised principal component analysis (PCA) revealed a clear separation between *GATA2* monoallelic and biallelic samples. Notably, the biallelic group (green) appears to cluster together even though it includes samples from two different individuals (IV.10_yr.4 and III.7). Moreover, a supervised gene expression signature analysis identified more than 2000 significantly differentially expressed (DE) genes (the large number of genes is likely a reflection of the small number of cases analysed) between these two groups, each driving a unique signature pattern (**Figure 3.6-B** and **Appendix 3. Table 2**). The significant DE genes were selected based on a stringent false discovery rate (FDR) filter of <0.05 and an absolute $\log_2$ fold change

(FC) value of >1. Precisely, 1,148 of these DE genes were downregulated and 1,284 were upregulated in *GATA2* biallelic compared to the monoallelic group (**Figure 3.6-C**).

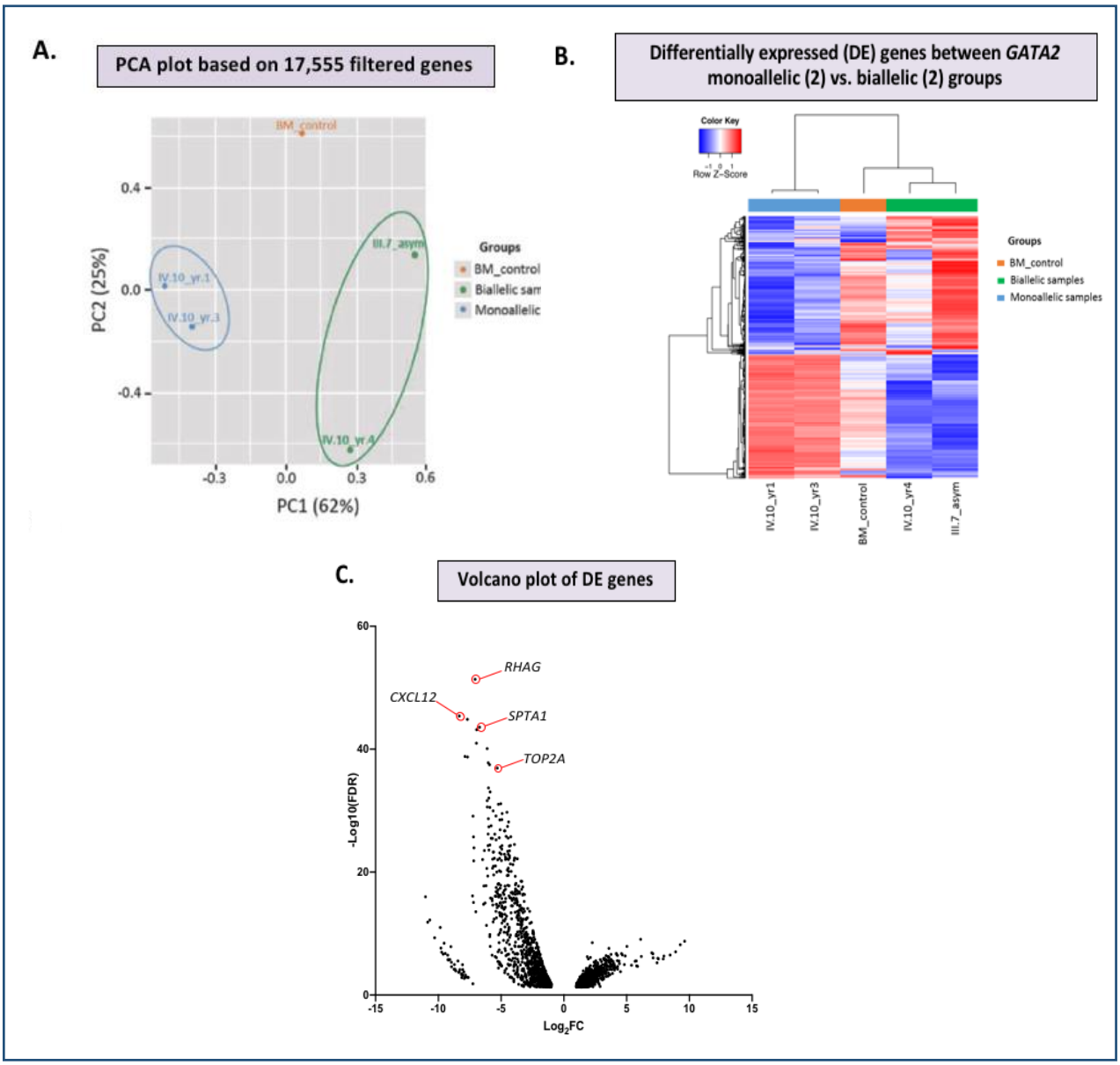


Figure 3.6 RNA-seq analysis. A. PCA plot showing a good separation between *GATA2* biallelic (green), monoallelic (blue) and BM control (orange) groups, based on all 17,555 filtered genes. **B.** Heat-map indicating 2,432 significantly DE genes between *GATA2* biallelic (green) and monoallelic (blue) samples, selected based on FDR value of <0.05 and an absolute log_2FC value of >1 (select DE genes are listed in **Appendix 3. Table 2**). **C.** Volcano plot indicating differential gene expression between these two groups and highlighting in red some of the most significantly DE genes, all of which are upregulated in *GATA2* biallelic vs. monoallelic samples.

3.5.1 GSEA of differentially expressed genes

Gene set enrichment analysis (GSEA) of RNA-seq data was next performed providing an opportunity to gain an appreciation of these DE genes and to explore significantly altered pathways (Subramanian et al., 2005). It was noteworthy that certain canonical pathways and oncogenic signature gene sets related to leukaemogenesis (e.g. cell cycle, DNA replication and HOXA9 signatures) were significantly downregulated (FDR <0.05) in *GATA2* biallelic compared to monoallelic samples as determined by their decreased normalised enrichment scores (NES), possibly explaining the oncogenic events and symptom manifestations observed in *GATA2* monoallelic samples (**Figure 3.7-A**). At the same time, a number of other pathways and gene sets implicated in regulating gene transcription, immune modulation and cell signaling including "RNA Pol 1 promoter opening", "Innate immune system" and "KRAS signatures" were found to be significantly upregulated (FDR <0.05) in *GATA2* biallelic vs. monoallelic samples, potentially reflecting improvements in clinical parameters and symptom recovery in *GATA2* biallelic samples which require further validation (**Figure 3.7-A**).

Furthermore, genes with E2F1 (cell cycle regulator) and GATA1 (which replaces GATA2 in erythroid and megakaryocytic differentiation) binding motifs in their regulatory regions were each shown to be significantly downregulated in *GATA2* biallelic vs. monoallelic groups whereas an enrichment of genes with CEBPA (GATA2 regulator) motifs was noted (although significance was not reached) (**Figure 3.7-B**). Of perhaps a greater significance, we observed that genes enriched for *GATA2* co-factor *PU.1* motifs in their regulatory regions were significantly overexpressed in *GATA2* biallelic vs. monoallelic samples (NES = 2.06) (**Figure 3.7-B**). This is in line with a recent finding by (Chong et al., 2017) who showed that p.T354M mutants bind to PU.1 with more affinity than WT GATA2 protein, thus resulting in sequestration of PU.1 from its

normal cellular functions (e.g. HSC survival and differentiation) and blocking its activity. As a result, the transcriptional activation brought about by PU.1 will likely be diminished in our *GATA2* monoallelic samples, where only the mutant allele is being expressed.

Taken together, these transcript profiles highlight the impact of losing expression of the WT allele (whilst still maintaining expression of the mutant allele) on *GATA2* activity and that of other haematopoietic genes downstream and how this allelic imbalance can lead to overt disease symptoms.

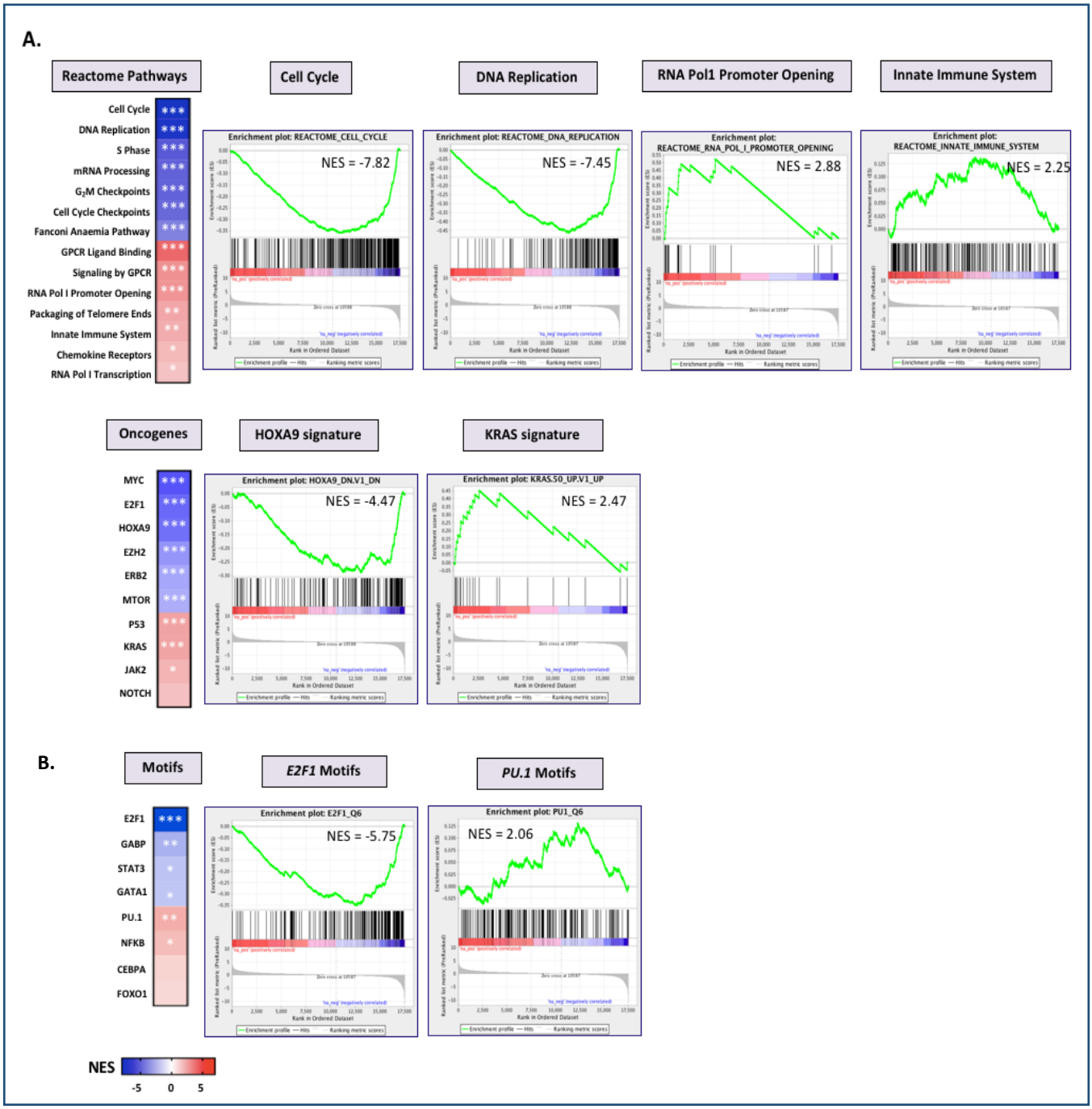


Figure 3.7 GSEA analysis of DE genes. Heat-maps indicating normalised enrichment scores (NES) from GSEA pathway analysis of RNA-seq data for **A.** reactome pathways and oncogenic signature gene sets and **B.** TF binding motifs (select GSEA plots are shown). A negative NES (**blue**) indicates a downregulation in *GATA2* biallelic vs. monoallelic samples whereas an opposite is the case for a positive NES (**red**). Statistical significance based on FDR q-values was determined at *p≤0.05, **p≤0.01 and ***p≤0.001.

3.6 Discussion

While most MDS/AML cases occur sporadically, there are rare occurrences of familial cases representing a high risk group of patients who require tailored management and optimal care. These rare patient populations are characterised by wide variations in the age of onset, disease latency and outcome, with many mutation carriers remain asymptomatic into late adulthood, making their identification, follow-up and management all the more challenging. The increased frequency of asymptomatic carriers due to reduced penetrance appears to be a feature of germline *GATA2* mutations, particularly in families segregating missense mutations such as p.Thr354Met. We believe that this observation opens an important question on leukaemia predisposition and how disease symptoms are initiated and potentially reversed. Since no published studies to date have addressed these questions in familial leukaemia, this study therefore sets out to uncover the molecular mechanisms governing the reduced penetrance and clinical heterogeneity in *GATA2* deficiency syndromes. Elucidating these mechanisms would have important clinical implications for counselling and management of this group of at-risk individuals.

The advent of NGS targeted gene panels has enabled us to decipher the pattern of secondary mutation acquisition at a greater depth sufficient to reveal haematopoietic subclones. Given that overt leukaemia typically in many cases reflects the acquisition of additional secondary mutations, targeted deep sequencing was employed across the symptomatic and asymptomatic family members in this study with a view of explaining the observed clinical heterogeneity that accompanies disease evolution from identical lesions. Our sequencing results suggest that the presence of somatic *ASXL1* mutations (p.Gly646TrpfsTer12) in the symptomatic patient and not in asymptomatic family members correlated with disease symptoms, corroborating a pathogenic

link between *GATA2* and *ASXL1*. This interpretation is in line with a report by West and colleagues (2014b) who noted molecular co-occurrence between germline *GATA2* and acquired *ASXL1* mutations in 14/48 (29%) of *GATA2*-deficient patients, with p.Gly646TrpfsTer12 *ASXL1* mutation (the one present in our family) being the most frequent by a significant margin (5/14).

ASXL1 is an epigenetic regulator involved in PRC2-mediated gene repression (Abdel-Wahab et al., 2012). LoF *ASXL1* aberrations were identified in sporadic AML and MDS cases (10% and 20% respectively), with a higher frequency in MDS/MPN cases such as chronic myelomonocytic leukaemia (CMML) (50%) and are associated with an adverse prognostic outcome (Boultwood et al., 2010, Asada and Kitamura, 2018). In addition, *ASXL1* mutations are recurrently found in healthy individuals with clonal haematopoiesis with indeterminate potential (CHIP) along with *DNMT3A* and *TET2* mutations (Genovese et al., 2014). Other examples where co-operating lesions ensue include 2nd biallelic *CEBPA* mutations in germline *CEBPA*-mutated cases (Tawana et al., 2015) and somatic *JAK2* or *CDC25C* mutations in germline *RUNX1*-mutated families (Yoshimi et al., 2014, Tawana et al., 2017b). Overall, these observations indicate that progression to overt malignancy requires the involvement of second “hits” to exert phenotypic effects, although in our symptomatic patient example (IV.10), *ASXL1* mutation VAF levels were low and maintained at (6-12%) over a 6-year monitoring period where spontaneous symptom recovery was noted at later time-points, suggesting that while these secondary mutations are important and in contrast to the assertion by West and colleagues, they are not sufficient alone to give rise to overt MDS or indeed determine when treatment is indicated.

Given that disease origin in *GATA2*-mutated patients is typically associated with haematopoietic stem cell exhaustion, recurrent infections and a resultant cytopenia, we would not discount the importance of monosomy 7, which was detected in the two cousins (IV.1 and IV.6) who developed MDS/AML, as a key secondary chromosomal anomaly. The aetiology for the selective loss of chromosome 7 in these patients remains unclear although the co-occurrence of monosomy 7 and LoF somatic *ASXL1* mutations in a *GATA2*-deficient background has been described in several reports (see **Table 3.1** below) but the order of these events has not been elucidated, reflecting perhaps the absence of longitudinal samples for analysis. However, on the basis of our observations and in agreement with previous studies (Wang et al., 2015, Pastor et al., 2017), it seems that monosomy 7 in IV.1 and IV.6 appears later than *ASXL1* mutations, therefore contributing to the malignancy but not initiating the cytopenia.

Table 3.1 The co-occurrence between *ASXL1* mutations and Monosomy 7 in germline *GATA2*-mutated families.

Family	Germline *GATA2* mutation	Acquired *ASXL1* mutation	Cytogenetic alteration	Reference
1	p.Asn371Lys	p.Leu817fs	Monosomy 7	West et al., 2014
2	p.Arg337X	p.Glu635fs	Monosomy 7	West et al., 2014
3	c.1018-2A>G, splice site	p.G646WfsX12	Monosomy 7	Wang et al., 2015
4	p.Gly200ValfsX18	p.G646WfsX12	Monosomy 7	Pastor et al., 2017
5	p.Ser447Arg	p.Thr655AsnfsX3	Monosomy 7	Pastor et al., 2017

The second key observation from this study was the reduced total *GATA2* expression levels in our symptomatic carrier (IV.10) accompanied by monoallelic expression favouring the mutant allele. In contrast, the higher *GATA2* expression in the asymptomatic carriers (III.5 and III.7) provided a protective buffer which substitutes function of the faulty gene. Remarkably, this allelic *GATA2* expression is dynamic and correlated with disease state in the symptomatic patient: (monoallelic expression – monocytopenia) in yr.1-3 and (biallelic expression – normal monocyte counts) in yr.4-6, providing a reasonable evidence of genotype-phenotype

correlations. This skewed *GATA2* expression is echoed further by a study in which MonoMAC patients exhibited reduced or absent expression of one *GATA2* allele due to mutations in an intronic 5 enhancer element, leading to decreased *GATA2* transcript levels and haploinsufficiency (Hsu et al., 2013). Monoallelic expression is therefore a phenomenon not only restricted to disease but also classical examples include imprinting of autosomal genes, where either the maternal or paternal allele is silenced, and the random inactivation of X-chromosomes in females (Reinius and Sandberg, 2015). Our data, however, provides the first example of reduced penetrance of germline coding mutations arising due to inter- and intra-individual variations in allelic expression and provides another plausible explanation for the variability in clinical presentations observed in a family with identical predisposing germline mutations.

Intriguingly and to our knowledge, when we assembled information from all *GATA2*-mutated MDS/AML pedigrees with reduced penetrance, we noted that the reduced penetrance phenotype was most prominent (if not exclusive) to germline missense LoF mutations such as p.Thr354Met (other examples are shown in **Table 3.2**). Indeed, as previously suggested by Kazenwadel et al. (2012), *GATA2* missense mutations might retain partial or residual activity, either in relation with DNA binding or interaction with partner proteins, rendering them less disruptive to protein function than *GATA2* truncating or frameshift mutations. Therefore, we believe our findings are most likely to have a similar impact in families with missense *GATA2* mutations other than p.Thr354Met, as it seems that the silencing of the WT allele expression would lead to an almost complete depletion of protein activity, constituting the initial step needed for symptoms to develop.

Table 3.2 Examples of germline *GATA2*-mutated families with reduced penetrance.

Family	Germline *GATA2* mutation	Asymptomatic carriers	Haematological malignancy	Reference
1	p.Thr354Met	2	MDS/AML	Hahn et al., 2011
2	p.Thr354Met	1	MDS/AML	Hahn et al., 2011
3	p.Thr354Met	1	MDS/AML	Hahn et al., 2011
4	p.Thr354Met	3	MDS/AML	Bödör et al., 2012
5	p.Arg398Trp	1	MDS/AML	Dickinson et al., 2014
6	p.Arg398Gln	3	MDS/AML	Dickinson et al., 2014
7	p. Ser447Arg	2	MDS/AML	Mir et al., 2015
8	c.10171+572 (C>T), intronic	1	MDS/AML	Churpek et al., 2015
9	p.Thr354Met	1	-	Unpublished (German family)
10	c.1017+582(G>T), intronic	3	-	Unpublished (German family)
11	p.Cys349Ser	1	-	Unpublished (German family)

For completeness, it is worth mentioning here that while genetic anticipation, the phenomenon whereby younger generations present with more severe disease phenotypes than their older counterparts, has been well-documented in *RUNX1*, *TERC* and *TERT*-mutated families, *GATA2* has not been associated with this phenomenon and we do not believe this to be the case in our family. Ages of disease onset for the deceased MDS/AML patients in the first and second generations (I.2, II.2 and II.3) were 53, 53 and 24 years old respectively, whereas third generation asymptomatic carriers (III.1, III.5 and III.7) are now in their mid-late 60s.

We would be the first to recognise that mutant *GATA2* monoallelic expression observed in the symptomatic vs. asymptomatic p.T354M carriers in this family would be strengthened further if a similar mechanism could be demonstrated in any of the other pedigrees mentioned. Indeed, our inability to access material for analysis across multiple pedigrees highlights one of the challenges in familial leukaemia research where available samples are limited, studies are primarily performed on retrospective material and consequently, DNA material is gathered rather than RNA and cells for subsequent studies.

In our experiments, even with ongoing collaborations with other groups including Hamish Scott's in Australia and Marcin Wlodarski's in Germany, we were unable to obtain matched DNA-RNA material from these p.T354M *GATA2*-mutated pedigrees, especially from earlier disease time points where *GATA2* monoallelic expression is most likely to be observed. We see ourselves as being rather fortunate therefore to have the case example (IV.10) where we have acquired multiple sequential RNA samples through routine clinical follow-up which allowed us to capture disease evolution and *GATA2* expression at the very early stages of the symptomatic patient's disease development. In a full-blown MDS/AML case scenario however, as in the two deceased cousins (IV.1 and IV.6), the ASE imbalance may not be so critical to the maintenance of malignancy and as such may go unnoticed.

Altogether, reduced penetrance indicates that germline *GATA2* p.T354M mutations may not be enough to lead to overt malignancy and while both monosomy 7 and *ASXL1* mutations are critical and important secondary events (**Figure 3.8**), the work presented in this chapter suggests that silencing of the wild-type *GATA2* allele observed in our symptomatic patient, leading to an almost complete abrogation of GATA2 function, is perhaps a more critical initiating event and a required step at the very early phases of disease in order to drive a patient's initial symptoms and the acquisition of secondary genomic lesions. In the next section (**Chapter 4**), we will examine the molecular mechanisms underpinning *GATA2* ASE and this loss of the WT allele expression.

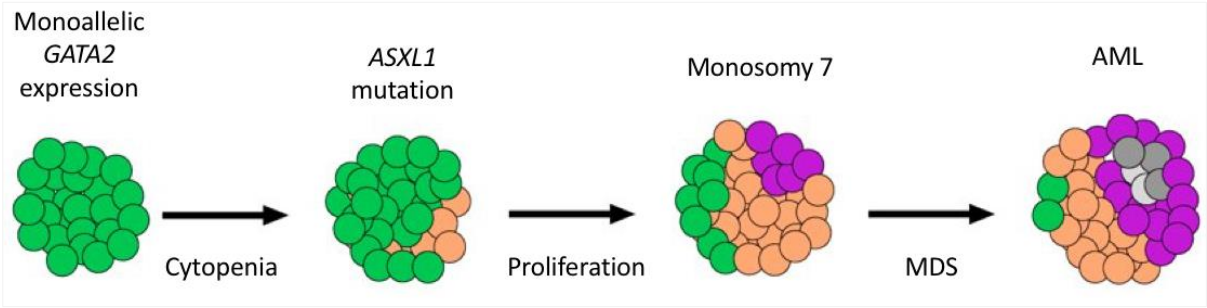


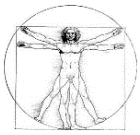
Figure 3.8 Model of clonal evolution of MDS/AML in a *GATA2*-deficient background.

Chapter 4. Results

Transcriptional Regulation of *GATA2* and Silencing of the WT Allele Expression

"To study cancer, is to also study its opposite. Before a cancer cell becomes corrupted by malignant genes, there is a code of normalcy. What does the normal (epi)genome do and how does it function? How does it ensure the astonishing diversity we have as humans and also the astonishing commonality?"

– Siddhartha Mukherjee



4.1 Background and Rationale

The experimental data in **Chapter 3** prompted us to explore the molecular mechanisms impacting on monoallelic *GATA2* expression. Reflecting for the most part the finite amount of primary material available to us, we focused our attention on non-coding genomic variants within *GATA2* regulatory regions and transient epigenetic mechanisms that could be responsible for the allele-specific changes in *GATA2* expression that accompany the p.T354M germline mutation in our symptomatic vs. asymptomatic carriers in the *GATA2*-mutated family. For the purpose of this thesis, the nitty-gritties of epigenetics and the mechanisms regulating gene expression are briefly reviewed in the following section.

4.2 Epigenetic Regulation of Gene Expression

In contrast with the focus on genetics in this thesis thus far, the term "epigenetics", first introduced by Conrad Waddington in the early 1940's, has evolved from the all-encompassing definition of "non-genetic heritable variations in phenotype" to now representing "the regulatory mechanisms that govern the expression of genetic information" (Waddington, 1942, Berger et al., 2009). Indeed, these mechanisms regulate gene expression through complex, dynamic, reversible and cell-specific post-translational modifications (PTMs) without changes in the DNA sequence itself (Holliday, 1987). Whilst many forms of epigenetic mechanisms do exist, two of the most frequently described (and briefly summarised in this section) are chemical alterations of the DNA (**DNA methylation**) and changes in DNA packaging via DNA binding proteins (**histone modifications**). These epigenetic marks decorate the DNA and its protein scaffold and, like music,

act as chemical notations, instructing the cells which genes to express and which ones to keep silent (Willyard, 2017).

4.2.1 DNA Methylation

DNA methylation involves the covalent addition of a methyl group to the 5'carbon of cytosine bases within CpG (Cytosine-phosphate-Guanine) dinucleotides (Bhalla, 2005). In cancer cells, this mainly takes place within CpG rich regions (~200 bases with GC content >50%) such as CpG islands located within promoter regions upstream of transcriptional start sites (TSSs) and typically constitutes a gene silencing mechanism by preventing the recruitment of TF machinery to facilitate gene transcription (Hashimshony et al., 2003). In somatic cells, however, these CpG islands are normally unmethylated (hypomethylated), whereas CpG sites outside of those regions are generally methylated (hypermethylated) (Weber et al., 2007) (**Figure 4.1**).

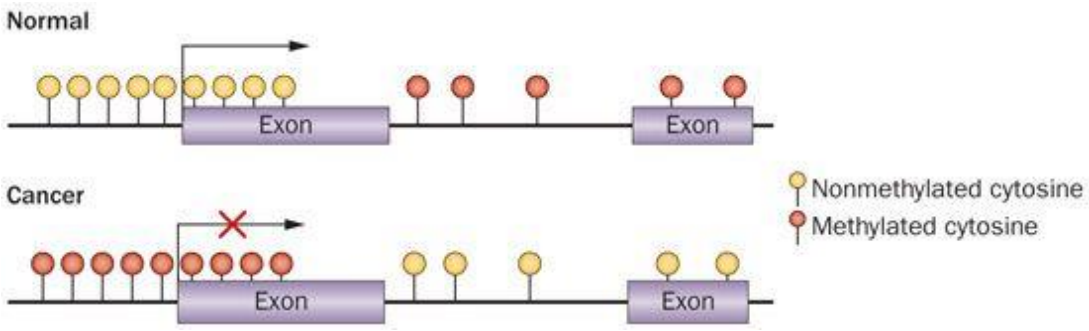


Figure 4.1 DNA methylation patterns in normal and malignant cells. Figure obtained from (Kandimalla et al., 2013).

Deposition of DNA methylation is catalysed by the DNA methyltransferase (DNMT) family of enzymes, which include DNMT1, DNMT3A and DNMT3B, all of which are essential for embryonic development (Okano et al., 1999, Holliday and Pugh, 1975). The removal of methyl groups from CpGs, however, is mediated by demethylase enzymes, namely TET1 and TET2, which oxidise 5-methylcytosine (5mC) to generate 5-hydroxymethylcytosine (5hmC) (Cortellino et al., 2011, He et al., 2011). Indeed, knockdown of Tet1 and Tet2 in mouse embryonic stem (ES) cells was shown

to downregulate genes associated with pluripotency and increase methylation at their promoters (Ficz et al., 2011). Of note, somatic *DNMT3A* and *TET2* are recurrently mutated in 10-20% of sporadic AML patients and confer an adverse impact on outcome (Delhommeau et al., 2009, Figueroa et al., 2010, Ley et al., 2010).

The role of DNA methylation in transcriptional regulation has been extensively studied. For example, it has been linked with X-chromosome inactivation in females, embryonic patterning, genomic imprinting and tissue-specific gene repression (Feinberg et al., 2002, Illingworth et al., 2008). In MDS/AML, promoter hypermethylation has been reported in *RUNX1* (Webber et al., 2013), *CEBPA* (Wouters et al., 2007, Lin et al., 2011) and *GATA2* (Shih et al., 2015, Celton et al., 2014). In particular, Celton and colleagues demonstrated a loss of *GATA2* expression in NK-AMLs and attributed this to aberrant DNA hypermethylation. Given that these PTMs are reversible, the use of demethylating agents such as 5-Azacytidine or 5-Aza-2-deoxycytidine has been shown to inhibit DNMTs and re-activate the expression of silenced genes (Leone et al., 2003).

More recently in relation to familial leukaemia, germline mutations in the DNA glycosylase (*MBD4*), which protects against methylation damage, were shown to predispose to inherited forms of MDS/AML, constituting a new candidate gene in familial disease (Sanders et al., 2018). DNA methylation also modulates disease penetrance; a study in monozygotic twins discordant for childhood leukaemia revealed that they displayed differential *BRCA1* methylation status (Galetzka et al., 2012).

Several techniques have been utilised for the analysis of DNA methylation including bisulphite sequencing. Sodium bisulphite treatment is considered the gold standard as it measures methylation at a single base resolution, converting unmethylated cytosines (C) into thymines (T)

whilst keeping methylated cytosines intact (**Figure 4.2**) (Frommer et al., 1992). This is then followed by either cloning for allele-specific quantitative approach (see methods **Chapter 2.7**) or deep sequencing of bisulphite-converted DNA for a more genome-wide analysis of DNA methylation (Lister et al., 2009).

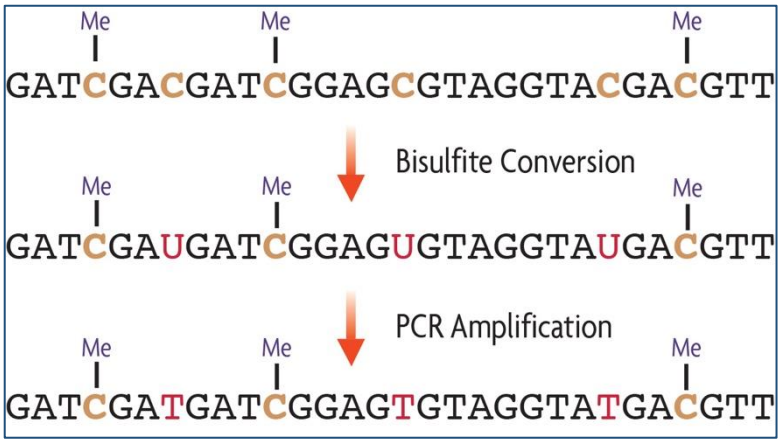
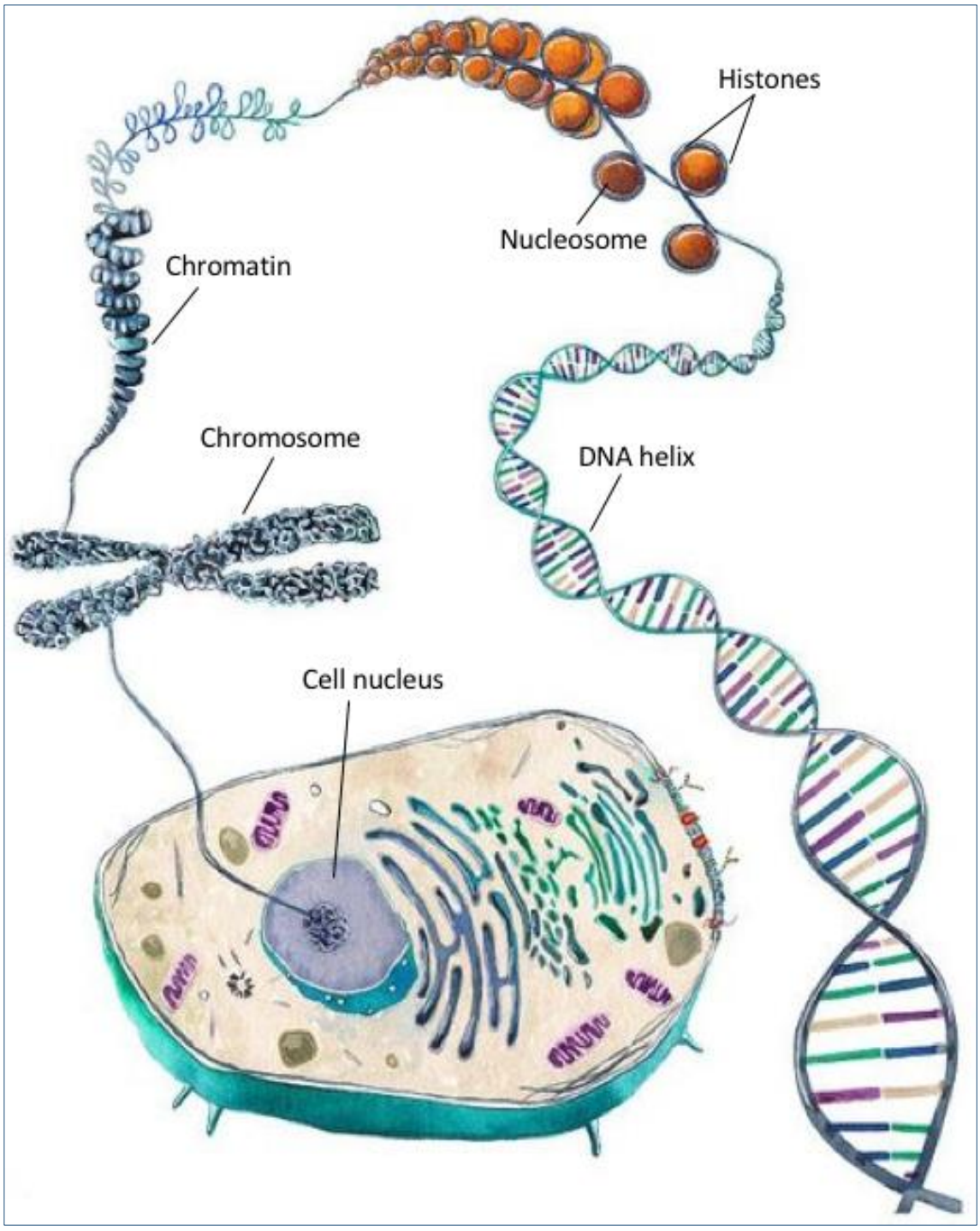


Figure 4.2 The effect of bisulphite DNA conversion on DNA sequence. Unmethylated cytosines are bisulphite-converted into uracils and upon PCR amplification, into thymines. Methylated cytosines remain unchanged following bisulphite conversion.

4.2.2 Chromatin Structure

In human cells, DNA is typically wrapped around octamers containing four histone proteins (H2A, H2B, H3 and H4) forming nucleosomes (Luger et al., 1997, Strahl and Allis, 2000). These nucleosomes are tightly packed into chromatins, akin to "beads-on-a-string", which are further coiled into chromatin fibers and then packaged into chromosomes (**Figure 4.3**). Each one of these highly conserved core histones has an N-terminal amino acid tail that protrudes outwardly, making it accessible for histone modifying enzymes to add specific PTMs to different amino acid residues (Kouzarides, 2007, Bannister and Kouzarides, 2011). These histone modifications orchestrate gene transcription, thereby facilitating the alternation of chromatin structure between two states; tightly compacted heterochromatin (**state 1**) which renders the DNA unattainable to TF machinery, whereas the opposite is the case for loosely packaged euchromatin (**state 2**), associated with transcriptionally active genomic regions (Henikoff, 2000, Bhalla, 2005).



<u>Figure 4.3</u> The structure and organisation of DNA packaging from the cell nucleus to the DNA double helix. Artwork obtained with permission from Sandra Black Culliton®.

4.2.3 Histone Modifications

Core histones have been shown to be enzymatically modified through methylation, acetylation, phosphorylation, ubiquitination, SUMOylation, ADP-ribosylation and arginine deamination, among others (Bannister and Kouzarides, 2011). Intriguingly, specific types of modifications only take place on particular amino acid residues. For instance, methylation is only found on arginine and lysine residues, the latter of which is the focus in this chapter (Zhang and Reinberg, 2001, Black et al., 2012, Dawson and Kouzarides, 2012).

These histone modifications collaborate together as part of a complex and dynamic "histone code", which regulates chromatin structure and thereby patterns of gene expression. Whilst deciphering the intricacies within this "histone code" is not an easy task, recent technological advances such as chromatin immunoprecipitation (ChIP) have helped immensely; it employs antibodies that specifically identify the histone modification of interest (Gade and Kalvakolanu, 2012). The ChIP-purified DNA can then be subjected to PCR and Sanger sequencing for an allelic-specific approach, utilising heterozygous variants at regulatory (promoter and/or enhancer) regions to distinguish between alleles, or deep sequencing (ChIP-Seq) for a more global analysis of TF binding sites (Dawson and Kouzarides, 2012, Barski et al., 2007).

4.2.3.1 Lysine Methylation

The majority of lysine methylation takes place on core histone 3 (H3). Frequently methylated residues include K4, K9, K27 and K36 among others, where they can be mono-(me1), di-(me2) or tri-(me3) methylated. Lysine methylation is a reversible process regulated by the opposing actions of lysine methyltransferases (KMTs) and lysine demethylases (KDMs) which respectively add and remove these methyl groups (Tian et al., 2013).

Lysine methylation can be linked with both transcriptional activation such as H3K4, H3K36 and H3K79 methylations, and repression such as H3K9 and H3K27 methylations, in a context-dependent fashion (Black et al., 2012). Herein, we focused on H3K4me3 and H3K27me3 to see if the enrichment of these chromatin marks can play a role in the ASE of *GATA2* in our patient samples.

Briefly, H3K4me3 is associated with transcriptionally active promoters and TSSs of transcribed genes and is mediated by the mixed-lineage leukaemia (MLL2) complex (Barski et al., 2007). Additionally, H3K4me3 was found to be inversely correlated with DNA methylation at CpG-rich promoters (Rose and Klose, 2014). H3K27me3, on the other hand, is catalysed by the polycomb repressive complex 2 (PRC2) including the histone methyltransferase EZH2 and is found at the promoters of repressed genes (Hock, 2012). Remarkably, while H3K27me3 and H3K4me3 seemingly exert opposite effects on chromatin activity, they can co-exist together as "bivalent chromatin marks" in promoters of TFs regulating lineage commitment in pluripotent stem cells, keeping them in a poised state for later activation during differentiation (Bernstein et al., 2006) (**Figure 4.4**).

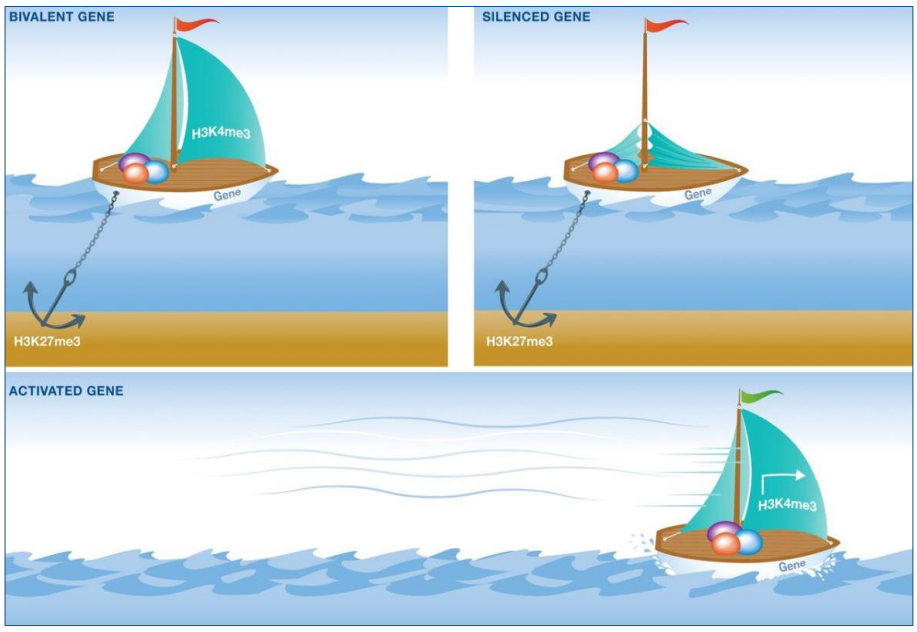


Figure 4.4 H3K4me3 (sail) and H3K27me3 (anchor) bivalent chromatin marks – an analogy. Figure obtained from (Harikumar and Meshorer, 2015).

Results

4.3 Variants within *GATA2* promoter regions were identified in the symptomatic but not in asymptomatic family members

There are many different potential mechanisms to explain ASE. In the following sections we investigate several of these in turn. First, to determine whether *cis*-regulatory variants within the *GATA2* locus could affect the binding affinity of TFs and therefore be implicated in the observed ASE, select members (III.5, III.7, IV.9 and IV.10) of the *GATA2* p.T354M-mutated family (described in **Chapter 3.2**) were screened for variants across key *GATA2* regulatory regions as defined by Ensembl and University of California Santa Cruz (UCSC) genome browsers and described in published literature e.g. (Pan et al., 2000, Hsu et al., 2013). We called these regions promoter 1 (GRCh38:3:128492893:128493893 – 1090bp), promoter 2 (GRCh38:3:128487621:128488621 – 1272bp) which also includes the 5'-UTR of *GATA2*, enhancer 1 (GRCh38: 3:128481374-128482164 – 217bp) located within an intronic region +9.5kb downstream of the TSS, and enhancer 2 (GRCh38:3:128601086:128603086 – 1655bp) which corresponds to -77kb region upstream of the TSS (**Figure 4.5-A**), representing 4234bp in total.

While Sanger sequencing detected no mutations in either of these 4 regions, two SNPs (rs9851497 [**T**/**C**] and rs1806462 [**C**/**A**]) located within *GATA2* promoters 1 and 2, respectively, were identified in the symptomatic (IV.10) but not in asymptomatic (III.5 and III.7) nor in WT (IV.9) family members, defining a distinct genotype between these two groups of mutation carriers (**Figure 4.5-B**).

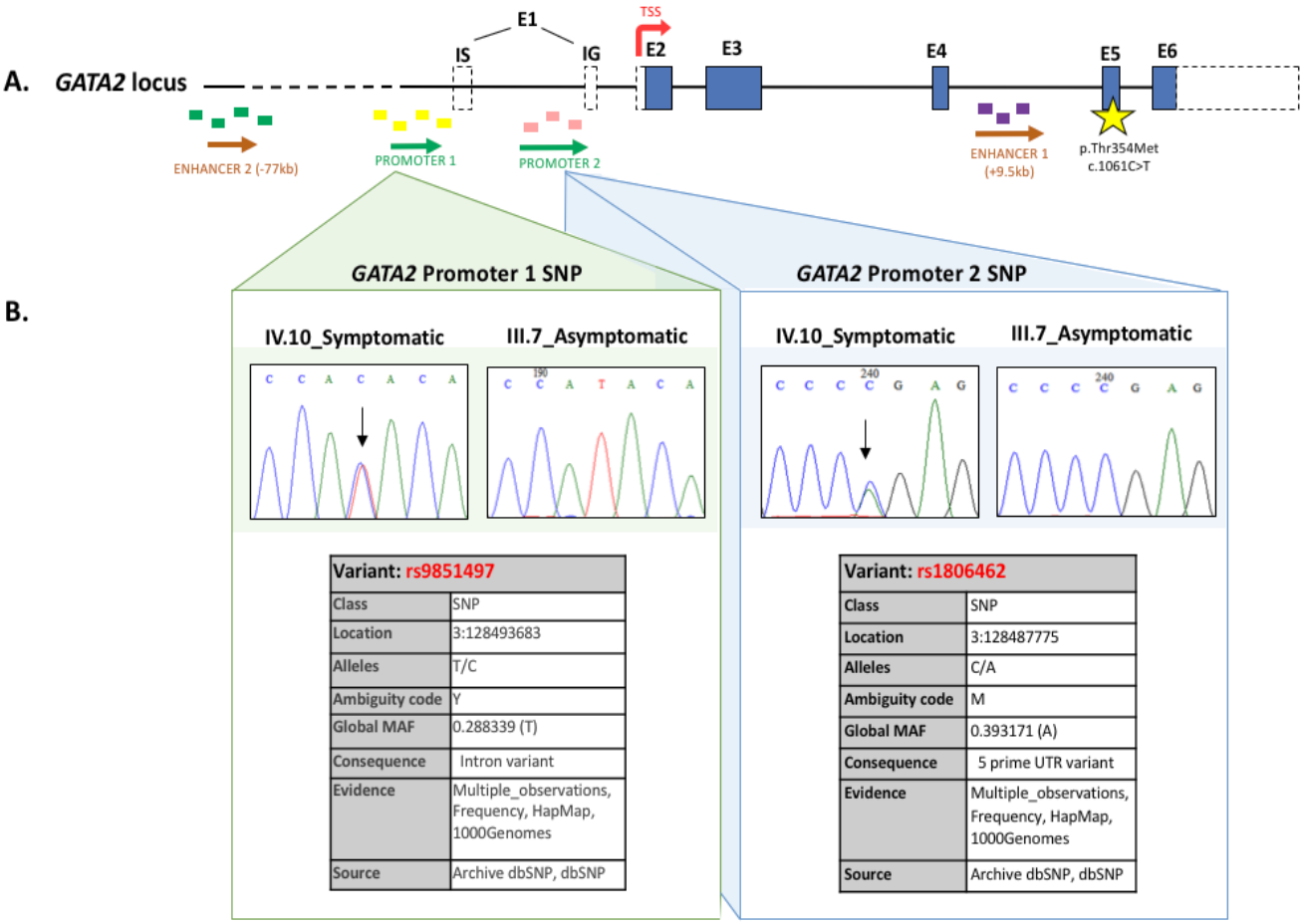


Figure 4.5 ***GATA2*** **regulatory variant screening. A.** A schematic representation of the *GATA2* locus (located on chromosome 3 – canonical transcript ENST00000341105.6) encompassing 6 exons and 2 promoter and 2 enhancer regions, with promoters 1 and 2 and enhancer 2 (-77kb) located upstream of the transcriptional start site (TSS) (red arrow) while enhancer 1 (+9.5kb) is located in an intronic region further downstream. The first exon of *GATA2* consists of 2 regions: the distal region (IS) is specific to the haematopoietic progenitor stem cell fraction and neuron, while the proximal region (IG) is found in tissues where GATA2 is typically expressed (Pan et al., 2000). The yellow star in exon 5 indicates the location of the germline *GATA2* mutation (c.1061C>T, p.T354M). The different coloured squares denote the locations of the overlapping PCR primers designed and used for variant screening (see methods **Chapter 2** - **Table 2.1**). **B.** Sanger sequencing chromatograms showing promoter 1 and 2 SNPs identified in the symptomatic (IV.10) vs. an asymptomatic (III.7) carrier together with variant details obtained from Ensembl and dbSNP (Database of Single Nucleotide Polymorphisms).

Notably, as promoter 2 SNP [**C**/**A**] (rs1806462) resides within the 5'UTR part of the gene, by sequencing the cDNA template of our symptomatic patient (IV.10), we were able to pinpoint/apportion which allele of the promoter 2 SNP (**A**) is allelic with the mutant allele (**T**) of the germline *GATA2* mutation (p.T354M, c.1061C>T) (**Figure 4.6**). This allowed us to attribute a distinct haplotype to the mutated allele and provided an opportunity to search for allele-specific molecular mechanisms governing the observed ASE of *GATA2* and a means to test our next hypotheses.

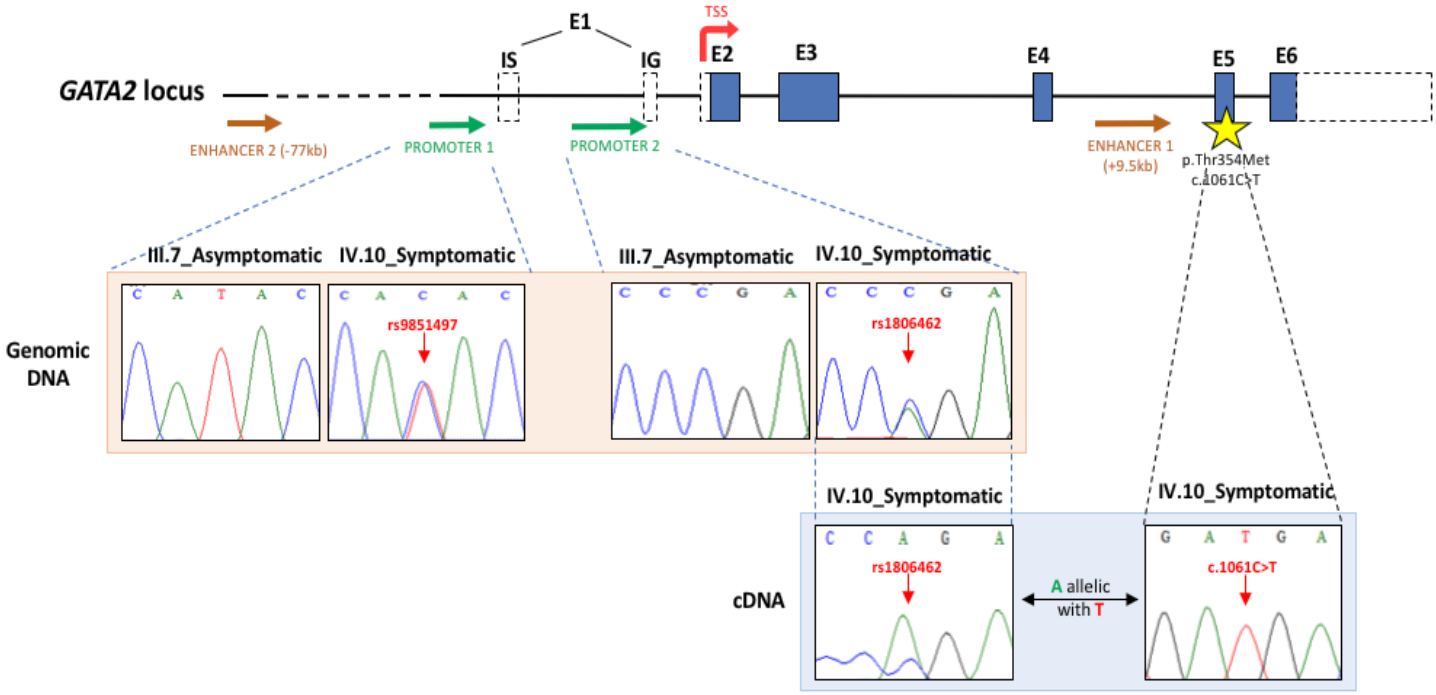


Figure 4.6 Establishing a haplotype on which the mutant allele lies. As promoter 2 is located within the 5'UTR of *GATA2* where it is transcribed, cDNA sequencing of the symptomatic carrier (IV.10_yr.1) showed that promoter 2 SNP allele (A), the one that is expressed, resides on the *GATA2* mutant allele (T) (c.1061C>T).

4.3.1 *GATA2* regulatory SNPs do not induce significant allele-specific differences in *GATA2* promoter activity

The identification of variants within *GATA2* promoter regions in the symptomatic patient prompted an investigation into the functional impact of these variants and how they contribute to ASE. To test whether there are any allele-specific differences in *GATA2* promoter activity due to the presence of these SNPs, luciferase reporter assay was performed; each allele of *GATA2* promoter 1 [**T**/**C**] (1318 bp) and promoter 2 [**C**/**A**] (1350bp) SNPs was amplified directly from our symptomatic patient (IV.10) (4 plasmid clones) and subcloned separately into a pGL2® luciferase reporter vector upstream of the Firefly luciferase gene (*Luc*) (Promega) (**Figure 4.7-A** and **B**). These constructs were then transiently co-transfected along with a pRL-CMV® control reporter vector containing Renilla luciferase gene (Promega) into HeLa cells. The luciferase activity of each promoter allele was next measured 48 hours post HeLa cell transfection, where the intensity of the luminescence from the Firefly luciferase was normalised to the luminescence intensity from the Renilla luciferase control after the addition of their respective substrates (see methods **Chapter 2.6** for a brief description of Firefly and Renilla luciferase chemistry).

Overall, while no major differences in luciferase activity were noted between promoter 1 SNP alleles [**T**/**C**], we observed a reproducible increase in the luciferase activity of the (**A**) allele compared with the (**C**) allele in promoter 2 SNP [**C**/**A**] (**Figure 4.7-C**). This increased transcriptional activity further reinforces the previously established link between the promoter 2 SNP allele (**A**) and the germline *GATA2* mutant allele (**T**) (the one solely expressed in IV.10 monoallelic samples), residing on the same haplotype (**Figure 4.6**).

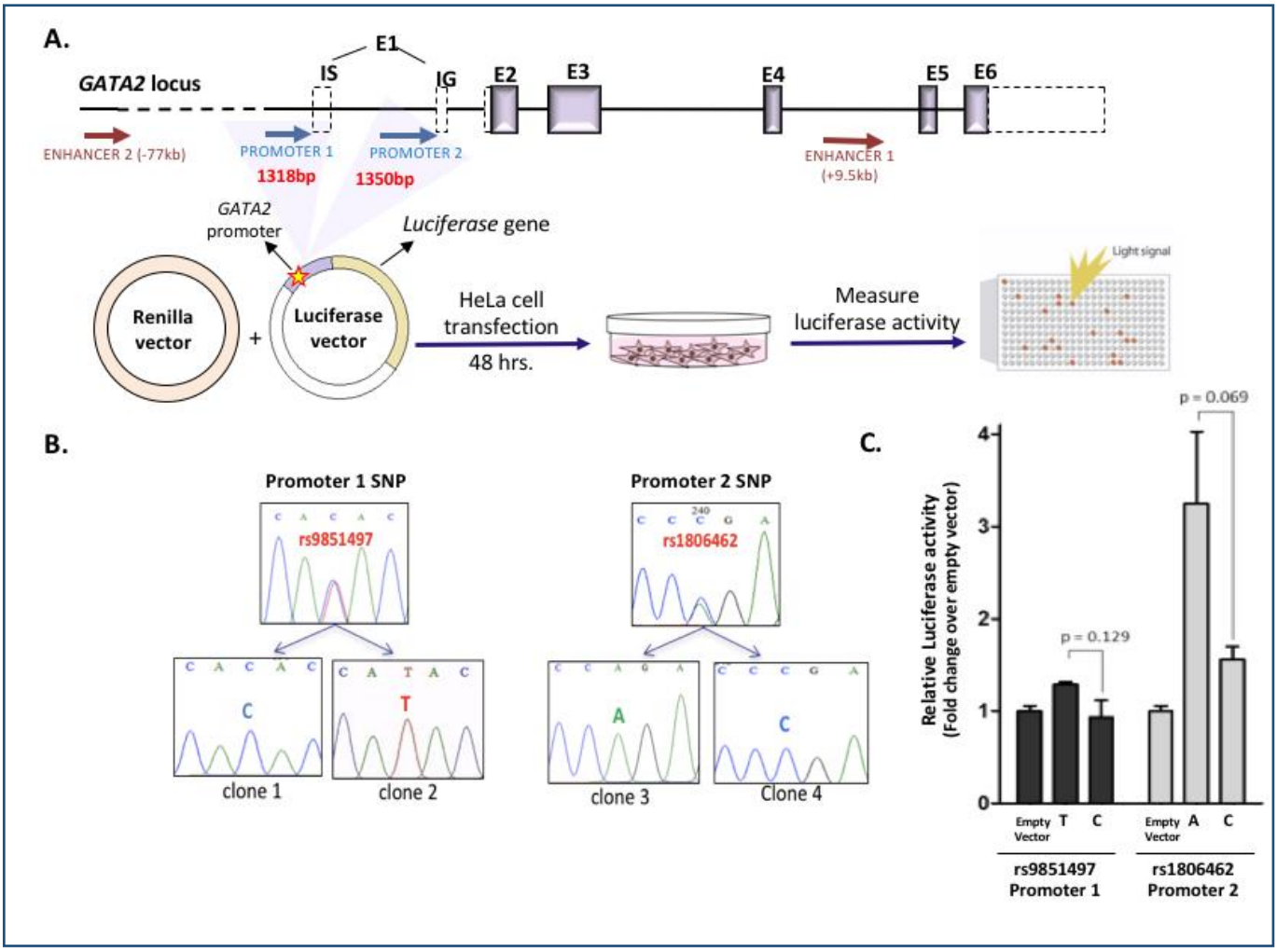


Figure 4.7 Functional validation of *GATA2* promoter SNPs. A. Workflow of the experiment procedure showing the Firefly luciferase reporter construct used to evaluate *GATA2* promoter activity. **B.** Sanger sequencing chromatograms demonstrating promoter 1 [T/C] and 2 [C/A] SNP allele clones that were obtained from IV.10 DNA following cloning of a 1318bp and 1350bp region (corresponding to *GATA2* promoters 1 and 2 respectively) which were then subcloned separately into a Firefly luciferase vector prior to HeLa cell transfection. **C.** Bar chart depicting relative luciferase activity of promoter 1 and promoter 2 alleles, calculated by normalising the intensity of Firefly luciferase to the intensity of Renilla luciferase luminescence and a final normalisation to an empty vector control. An average of 5 independent experiments is shown, including all samples in triplicates. Statistical significance was determined at *p ≤ 0.05, **p ≤ 0.01 and ***p ≤ 0.001. Error bars represent SEM.

The potential role of these promoter SNPs with regards to differential TF binding affinity was assessed using PROMO *in silico* prediction tool (Messeguer et al., 2002). As illustrated in **Figure 4.8**, a substitution from **C** to **T** in promoter 1 SNP [**T**/**C**] can result in the creation of new binding sites for TFs *XBP1* and *YY1* while no allele-specific differences in TF binding occupancy were noted in promoter 2 SNP [**C**/**A**]. This absence (*in silico*) of differential TF binding affinity in promoter 2 SNP and the transient nature of ASE have led us instead to explore the epigenetic control of *GATA2* allele-specific expression, described in the ensuing **sections** (**4.4** to **4.7**).

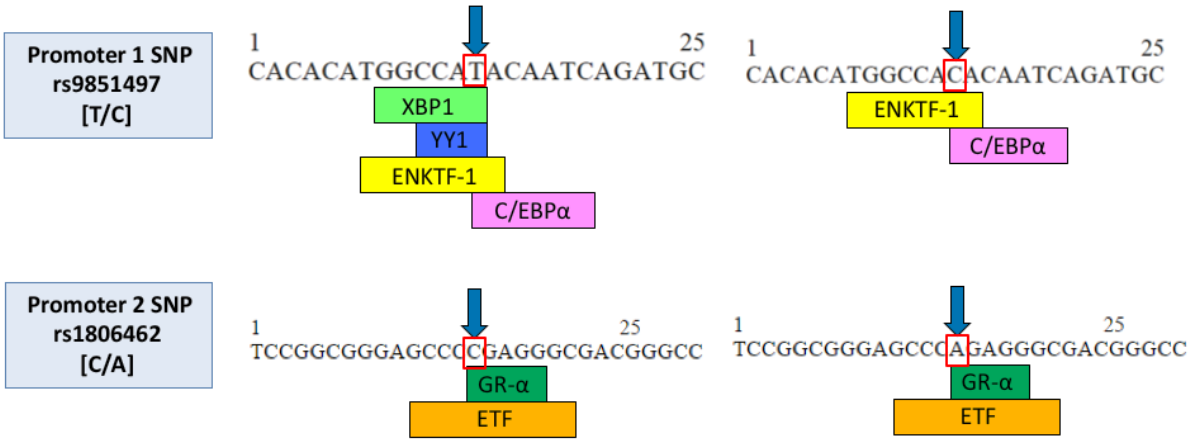


Figure 4.8 PROMO-based allele-specific TF binding site prediction for promoter 1 and 2 SNPs. Promoter 1 SNP demonstrates allele-specific differences in TF binding sites between alleles **T** and **C**, whereas it is not the case for promoter 2 SNP alleles **C** and **A**.

4.3.2 The second promoter CpG-SNP provided a means of distinguishing between alleles

Herein, we utilised the second promoter SNP (rs1806462 [**C**/**A**]) as a marker to differentiate between mutant and WT alleles in our subsequent experiments (**Figure 4.9-A**) taking advantage of the fact that this second promoter SNP is located within the 5'UTR (and as mentioned earlier), we were able to define a haplotype between the SNP allele (**A**) and the germline mutant *GATA2* allele (**T**) (**Figure 4.9-B**). Interestingly, we also noted that this SNP creates/abolishes a CpG dinucleotide within the *GATA2* promoter region. As shown in **Figure 4.9-C**, this SNP can remove a CpG methylation site within the mutant allele (**A**) and generate a new CpG methylation site within the WT allele (**C**). In subsequent experiments, and focusing on epigenetic regulation of *GATA2* second promotor region, we used this CpG-SNP as a vehicle to test whether allele-specific differences in DNA methylation and chromatin mark deposition provide an explanation for the silencing of the WT *GATA2* allele expression observed in the earlier time-points of IV.10 and whether this SNP has also a role in modulating (increasing or lowering) methylation at this region. Indeed, one would postulate that losing (or gaining) a CpG site due to the presence of these CpG-SNPs could exert an effect on the expression (or lack thereof) of mutant/WT alleles, such that the silencing of the WT allele can be attributed to the presence of extra CpG sites within the promoter region, coinciding with increased DNA methylation. Conversely, the expression of the mutant allele would be explained by the loss of CpG sites and a reduction in methylation overall.

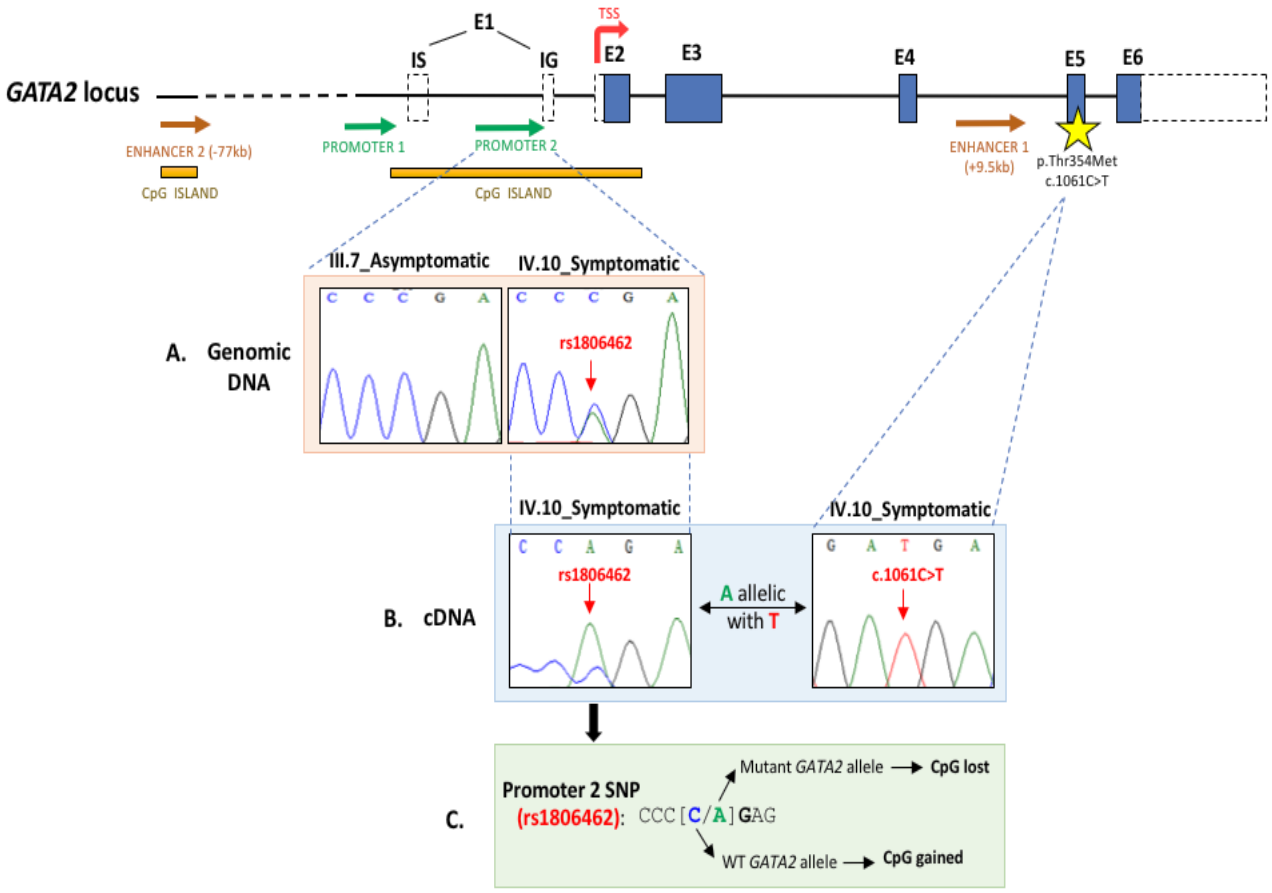


Figure 4.9 Focusing on *GATA2* second promoter SNP. A. This promoter SNP (rs1806462 [C/A]) overlaps a CpG island and was only present in the symptomatic patient (IV.10). It is situated 5,874 bp away from the p.T354M germline mutation site (denoted by a yellow star). **B.** Given the location of promoter 2 SNP within the 5'UTR, a haplotype between the SNP allele (A) and the germline mutant allele (T) was established by cDNA sequencing of IV.10_yr.1, providing a key to distinguish between mutant and WT alleles. **C.** This promoter SNP [C/A] also removes a CpG methylation site within the mutant allele (A) and generates an extra CpG methylation site within the WT allele (C).

4.4 No differences in global DNA methylation were observed between symptomatic and asymptomatic family members

Initial experiments focused on comparing *GATA2* methylation patterns between symptomatic and asymptomatic carriers within the family to establish if there is an increase in global DNA methylation of *GATA2* promoter 2 region leading to the observed *GATA2* transcriptional silencing in the symptomatic carrier (IV.10). Methylation-specific PCR (MSP) of bisulphite-modified DNA was performed using primers covering *GATA2* promoter 2 CpG islands that can specifically amplify methylated (M) and/or unmethylated (U) sequences (see **section 4.2.1** and methods **Chapter 2.7** for a summary of bisulphite DNA conversion). Based on the PCR gel in **Figure 4.10**, promoter 2 was shown to be methylated in the symptomatic carrier (IV.10). However, there were no major differences in methylation patterns (M and U) when compared with asymptomatic (III.5 and III.7) and WT (IV.8 and IV.9) family members.

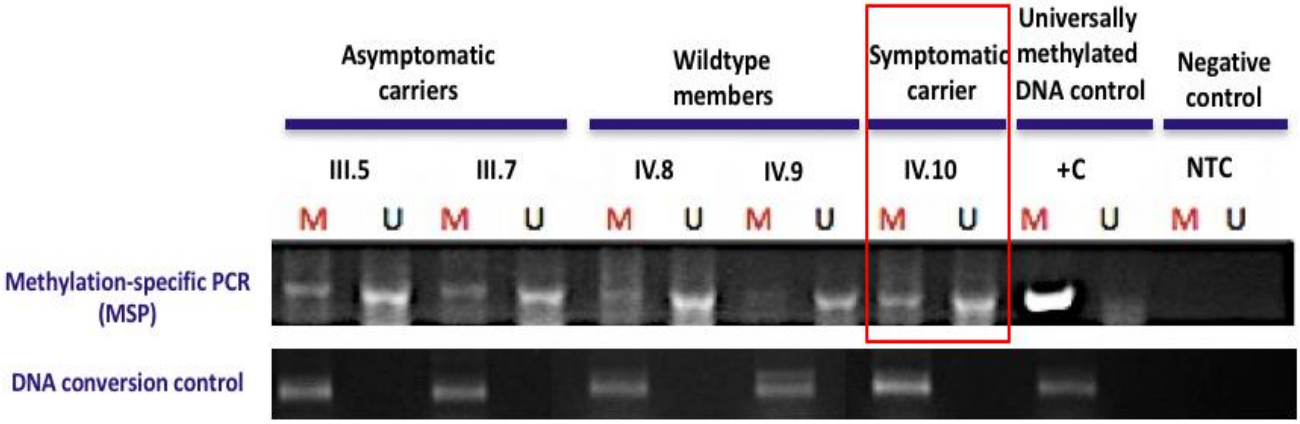


Figure 4.10 Global DNA methylation profiles. A representative MSP gel showing PCR products of bisulphite-modified DNA from select *GATA2*-mutated family members amplified using *GATA2* promoter 2 methylated (M) and unmethylated (U) primer set. Positive and negative controls are shown for comparison together with a bisulphite DNA conversion control in the second gel lane using conversion-specific primer set from Active Motif (see methods **Chapter 2.7** and **Table 2.1** for a list of primers used).

4.5 Allele-specific DNA methylation as a regulatory mechanism of silencing the WT *GATA2* allele expression

It is important to keep in mind that the assay used in the previous section (**4.4**) assesses global DNA methylation and does not specifically distinguish between mutant and WT alleles. We therefore utilised the CpG-SNP [**C**/**A**] residing within *GATA2* promoter 2 to determine whether allele-specific differences in DNA methylation is a contributing factor to *GATA2* ASE. Bisulphite-specific PCR (BSP) was performed to amplify a 200-bp region encompassing *GATA2* promoter 2 SNP and overlapping a CpG island (containing ~20 CpG sites) followed by cloning and sequencing, to compare DNA methylation patterns between mutant and WT alleles across three time-points of IV.10. (yr. 1, 3 and 6) (See methods **Chapter 2.7**).

As illustrated in **Figure 4.11-A-B** and **Figure 4.12**, we observed a significant increase in promoter methylation in the WT allele (**C**) compared to a reduced methylation pattern in the mutant allele (**A**) of IV.10 earlier time-points (yr. 1 and 3). Interestingly, the extra CpG site in the WT allele (**C**) generated by this CpG-SNP [**C**/**A**] was frequently methylated, probably favouring the increase in DNA methylation and subsequent ASE in these earlier time-points. Conversely, we observed no significant allele-specific differences in methylated CpGs in IV.10_yr.6, which offers a potential explanation for the restoration of the WT allele expression and an improvement in clinical parameters at this later time period.

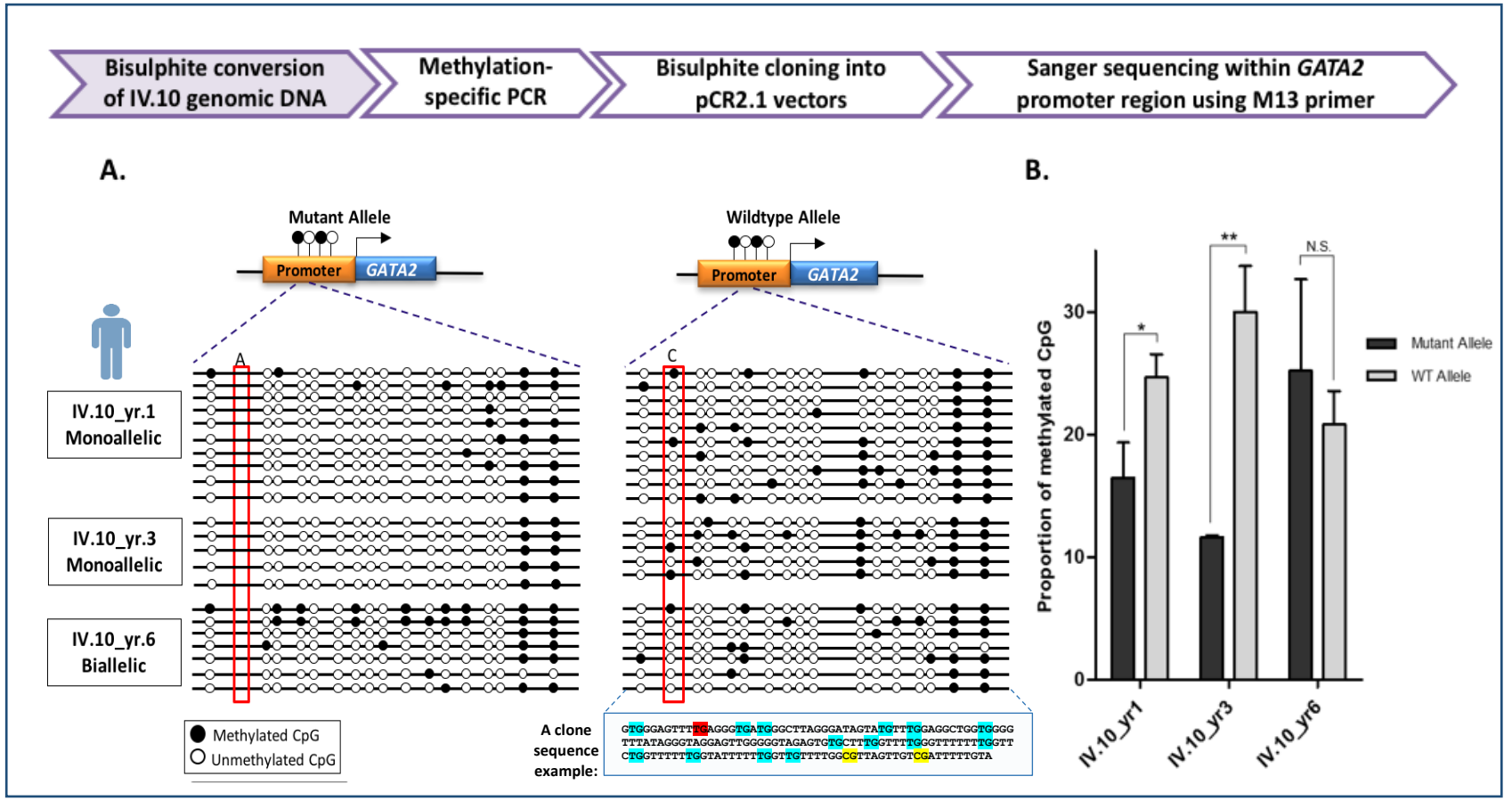


Figure 4.11 Allele-specific DNA methylation profiles. A. Bisulphite cloning and sequencing covering *GATA2* second promoter SNP region overlapping a CpG island was performed to compare DNA methylation patterns between mutant and WT alleles across 3 time-points of the symptomatic patient IV.10 (yr. 1 and 3 with monoallelic *GATA2* expression) and (yr.6 with biallelic *GATA2* expression). Each row represents a separate clone. Each circle represents a CpG site, black circles correspond to methylated CpGs while white circles correspond to unmethylated CpGs. **B.** The proportion of methylated CpGs between mutant and WT alleles across the 3 time-points studied. WT allele is significantly more methylated than the mutant allele in monoallelic samples (yr. 1 and 3), whereas no significant allele-specific differences in methylation were observed in a biallelic-expressing sample (yr. 6). An average of 3 independent experiments is shown. Statistical significance was determined at *p <0.05, **p <0.01, and ***p <0.001 using a *t*-test with Bonferroni correction. NS: non-significant comparisons. Error bars represent SEM.

Original DNA sequence:

GGCCTCTGAGAGTGAAGGAGTTCCGGCGGGAGCCCCGAGGGCGACGGGCCCAGGGACAGCACGTCCGGAGG
CTGGCGGGGCTTACAGGGTAGGAGCTGGGGGTAGAGTGCGCCTCGGCCTCGGGCCCGCCCGGCTCCGGCCC
CTCGGCATCCTCCGGCCGCCCTGGCGCCAGCTGCCGACTCCTGCACAGACATGAAGCGGGGGCCGCGCACG

Bisulphite-converted DNA sequence:

GGTTTTTGAGAGTGAAGGAGTTTCGGCGGGAGTTTCGAGGGCGACGGGTTTAGGGATAGTACGTTCGGAGG
TTGGCGGGGTTTATAGGGTAGGAGTTGGGGGTAGAGTGCGCTTCGGTTTCGGGTTCGTTCGGTTTCGGTTT
TTCGGTATTTTTCGGTCGTTTTGGCGTTAGTTGTCGATTTTTGTATAGACATGAAGCGGGGGCCGCGCACG

Key:

Total number of CGs = 18
GGTTTTTGAGAGTGAAGGAGTTTC = forward primer sequence
CTATACAAAAATCGACAACTAACGC = reverse primer sequence
CG = Promoter 2 SNP [C/A] removes a CpG site within the mutant allele (A) and creates a CpG site within the WT allele (C)
TG = unmethylated CpG
CG = methylated CpG

Representative examples of mutant and WT allele clones (as shown in Figure 4.11-A):

IV.10_yr.1 clone 1 (wildtype allele CG) TG = 13 CG = 5

TGAGAGTGAAGGAGTTTCGGTGGGAGTTTCGAGGGTGATGGGTTTAGGGATAGTATGTTCGGAGGTTGGTG
GGGTTTATAGGGTAGGAGTTGGGGGTAGAGTGTGTTTTGGTTTTGGGTTTGTTCGGTTTTGGTTTTTTGGT
ATTTTTTGGTTGTTTTTGGCGTTAGTTGTCGATTTTTTGTATAGA

IV.10_yr.1 clone 1 (mutant allele AG) TG = 13 CG = 4

TGAGAGTGAAGGAGTTTCGGCGGGAGTTTAGAGGGTGACGGGTTTAGGGATAGTATGTTTGGAGGTTGGTG
GGGTTTATAGGGTAGGAGTTGGGGGTAGAGTGTGTTTTGGTTTTGGGTTTGTTTGGTTTTGGTTTTTTGGT
ATTTTTTGGTTGTTTTTGGCGTTAGTTGTCGATTTTTTGTATAGA

Figure 4.12 Allele-specific DNA methylation analysis. Two representative examples of mutant (A) and WT (C) allele clones obtained from IV.10_yr.1 DNA following bisulphite cloning and sequencing encompassing *GATA2* promoter 2 SNP region (~200bp) and overlapping a CpG island (containing 18 CGs). Sequences highlighted in pink represent the primer pair used to amplify this region prior to cloning into pCR2.1® TA-vectors (see **Chapter 2.7** and **Table 2.1**). This approach was used to quantitatively measure the proportion of methylated (CG) and unmethylated (TG) CpGs across IV.10 different time-points and utilising CG promoter 2 CpG-SNP [C/A] as a way of distinguishing between mutant (A) and WT (C) alleles.

4.6 Elevated H3K4me3 promoter deposition on the mutant allele

Given the inverse correlation between DNA methylation and gene expression, we next sought to establish whether these allele-specific changes in *GATA2* methylation and expression are accompanied by changes in chromatin deposition at the promoter. We focused specifically on 2 histone marks, H3K4me3 and H3K27me3. As mentioned in section **4.2.3.1**, H3K4me3 (mark of gene activation) and H3K27me3 (mark of gene silencing) define poised or closed chromatins, respectively, rendering them more or less accessible for TFs, thereby representing potential regulators of gene expression.

To assess if these bivalent chromatin marks are differentially enriched between mutant and WT alleles of IV.10 samples, allele-specific ChIP was performed which briefly involved protein to DNA crosslinking, chromatin sonication and immunoprecipitation using either H3K4me3 or H3K27me3 antibodies followed by reverse cross-linking and PCR and Sanger sequencing within *GATA2* promoter 2 SNP region (see methods **Chapter 2.8**). Access to primary material was critical to these experiments as it allowed us to test three sequential BM samples from IV.10: two early monoallelic samples (yr. 1 and 3) and a later time-point with biallelic *GATA2* expression (yr.4), using the SNP (rs1806462 [**C**/**A**]) genotype as a means of discriminating between mutant and WT alleles.

As shown in **Figure 4.13-A** and **B**, while there were no apparent allele-specific differences in H3K27me3 inhibitory mark deposition across the different time-points of IV.10, we noted a significant enrichment in the deposition of H3K4me3 activating mark on the promoter mutant allele (**A**) relative to the WT allele (**C**) in IV.10 monoallelic samples (yr. 1 and 3). Critically and consistent with the pattern observed in DNA methylation, there were no demonstrable allele-specific differences in H3K4me3 deposition in the IV.10 biallelic sample at the later time point

(yr. 4) (both alleles equally enriched for H3K4me3) which coincided with apparent reactivation of the WT allele expression and a steady improvement in IV.10 disease symptoms.

PCR quality control assay was performed to confirm that these ChIP Sanger sequencing results were not due to bias introduced by PCR. Briefly, this entailed taking promoter 2 SNP [**C**/**A**] plasmid DNA containing either mutant (**A**) or WT (**C**) alleles that were cloned from IV.10 DNA (and were generated for the luciferase assay experiments in **section 4.3.1**) and introducing them together into PCR reactions at different ratios (1:1,2:1,3:1,4:1,6:1) of **A**:**C** and the reverse (1:1,1:2,1:3,1:4,1:6) of **A**:**C**. These ratios were measured and confirmed by Sanger sequencing trace chromatograms where the proportion of sequence peak height corresponding to each allele is compared against the expected ratios of added DNA (**Figure 4.14**).

Taken together, these findings support the view that monoallelic *GATA2* expression is driven in part by dynamic epigenetic reprogramming; increased DNA methylation linked with lower H3K4me3 promoter deposition on the WT allele and vice versa for the mutant allele, potentially underlying the phenotypic variation and disease penetrance of germline *GATA2* p.T354M mutations.

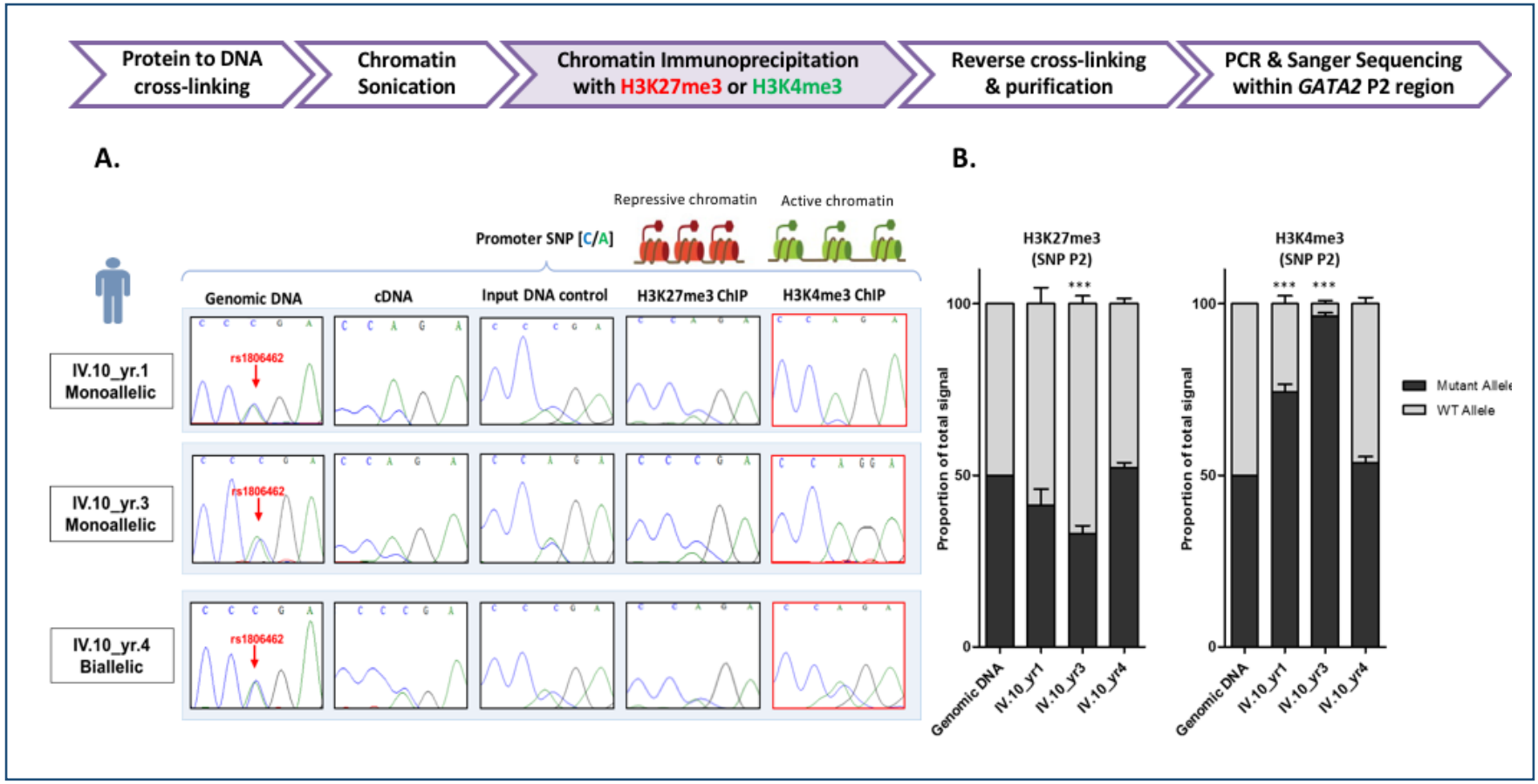


Figure 4.13 Allele-specific enrichment of H3K4me3 and H3K27me3 chromatin marks. A. The first 2 columns from the left (genomic DNA and cDNA) represent Sanger sequencing trace chromatograms obtained **before** ChIP and following PCR for a region spanning *GATA2* promoter SNP (rs1806462) [C/A] where (A) corresponds to the mutant allele and (C) corresponds to the WT allele. The remaining 3 columns represent Sanger sequencing traces obtained **after** ChIP and showing input DNA control, ChIP for H3K27me3 and H3K4me3, respectively, across three sequential time-points of the symptomatic patient (IV.10_yr.1 and yr.3) with monoallelic *GATA2* expression and (IV.10_yr.4) with biallelic *GATA2* expression. **B.** Quantification of mutant and WT allele ChIP sequence peak heights across the time-points of IV.10 based on Sanger sequencing. H3K4me3 activation mark favoring the mutant allele was enriched in monoallelic samples (yr. 1 and yr. 3) compared with the biallelic expressing sample (yr. 4) whereas no significant differences in H3K27me3 deposition were noted across the time-points. An average of 3 independent experiments is shown. Statistical significance was determined at *p <0.05, **p <0.01, and ***p <0.001 using a *t*-test with Bonferroni correction. Error bars represent SEM.

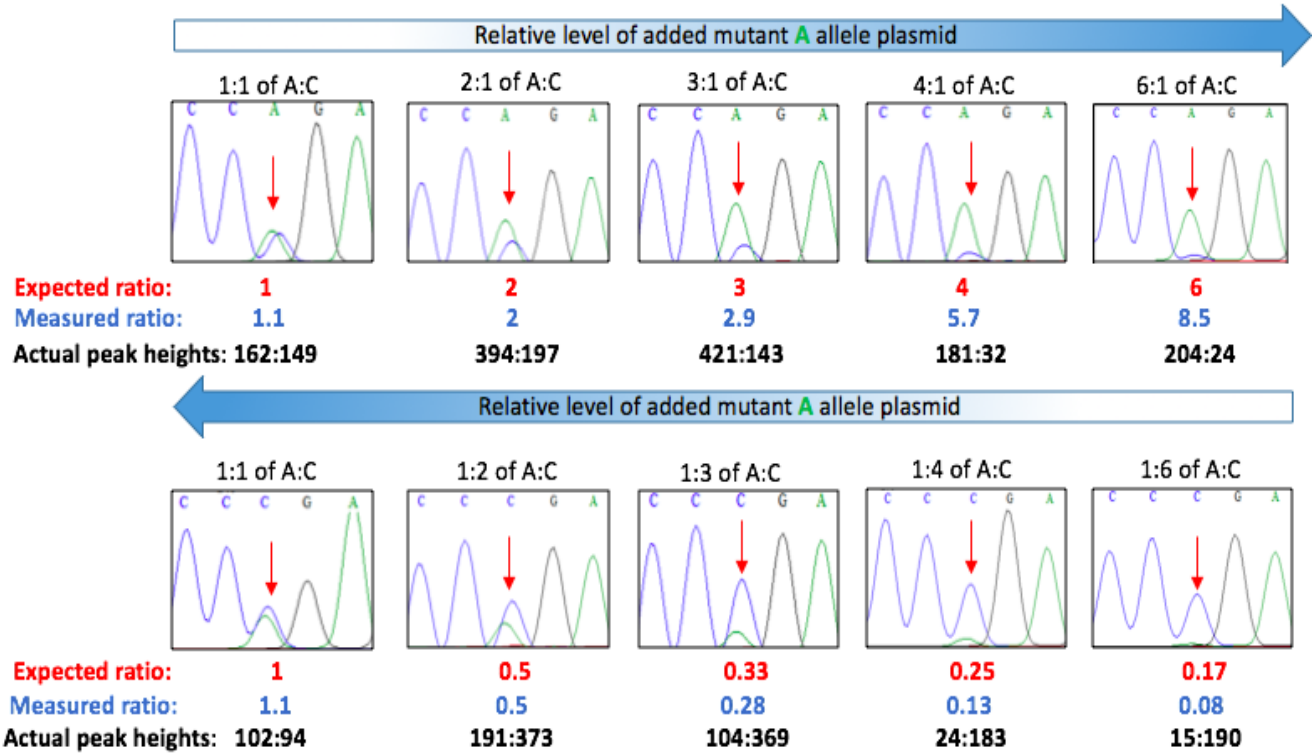


Figure 4.14 PCR quality control assay. This assay was performed to confirm that Sanger sequencing results were not due to bias introduced by PCR. Plasmids containing either mutant (A) or WT (C) promoter 2 SNP alleles [C/A] were introduced together in PCR reactions at different ratios (1:1,2:1,3:1,4:1,6:1) of A:C and the reverse (1:1,1:2,1:3,1:4,1:6) of A:C which was then confirmed by Sanger sequencing chromatograms covering the promoter SNP. Here, the proportions of sequence peak heights (measured ratios) are consistent with the expected ratios of added DNA.

4.7 Investigating the relationship between DNA methylation and chromatin mark deposition in our patient samples

4.7.1 H3K4me3 promoter deposition appears to be mutually exclusive with DNA methylation

In light of our previous observations that monoallelic *GATA2* expression is associated with allele-specific DNA methylation and H3K4me3 promotor deposition, we wanted to test whether these two events occur simultaneously or indeed if they are mutually exclusive at the *GATA2* locus within our patient samples. The lines of evidence that the latter scenario may hold true in mice and human cells are two-fold:

1) It has been demonstrated that H3K4me3 occupancy blocks *de novo* DNA methylation by preventing the recruitment of DNA methylation machinery e.g. Dnmt3L and Dnmt3a/b tetramers to promoters. These DNMT3 enzymes contain an ADD domain that specifically recognises unmethylated H3K4 but cannot bind to H3K4me3.

2) ZF-CxxC domain proteins such as the Mll1/2 complexes or Cfp1 recruit the Set1a/b H3K4-methyltransferase complexes to unmethylated DNA in CpG islands, where they function to catalyse H3K4me3 (Rose and Klose, 2014) (**Figure 4.15**).

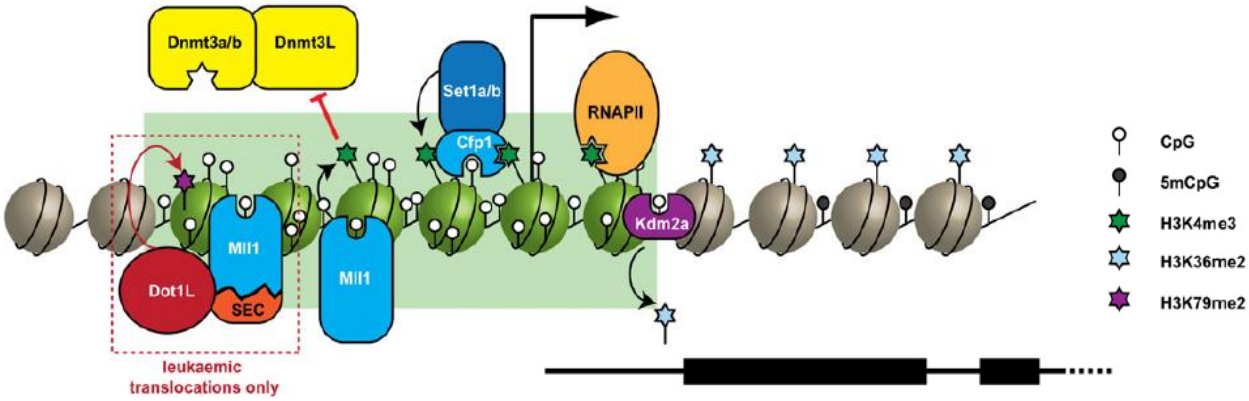


Figure 4.15 A schematic diagram depicting the associations between H3K4me3 and DNA binding proteins at CpG regions. CpG islands associated with actively transcribed genes recruit H3K4-methyltransferases (e.g. Set1a/b) through interactions between ZF-CxxC domains (e.g. Cfp1 and MII1/2) and unmethylated CpG regions. RNA PolII also associates with H3K4me3 at active gene promoters. H3K4me3 can also block DNA methylation by preventing the binding of Dnmt3L/Dnmt3a/b tetrameric complex to CpG islands. Figure adapted from (Rose and Klose, 2014).

To test this hypothesis in our patient samples, we performed BSP of bisulphite-modified, H3K4me3 ChIP-enriched DNA from the previous ChIP experiments followed by cloning and sequencing within *GATA2* promoter 2 region to assess DNA methylation status. As illustrated in **Figure 4.16**, this revealed little or no DNA methylation at any sampling time-point of IV.10's H3K4me3-enriched cells (yr.3 *GATA2* monoallelic and yr.4 *GATA2* biallelic expressing samples), providing direct experimental evidence that DNA methylation and H3K4me3 promoter deposition appear to be mutually exclusive in our IV.10 samples.

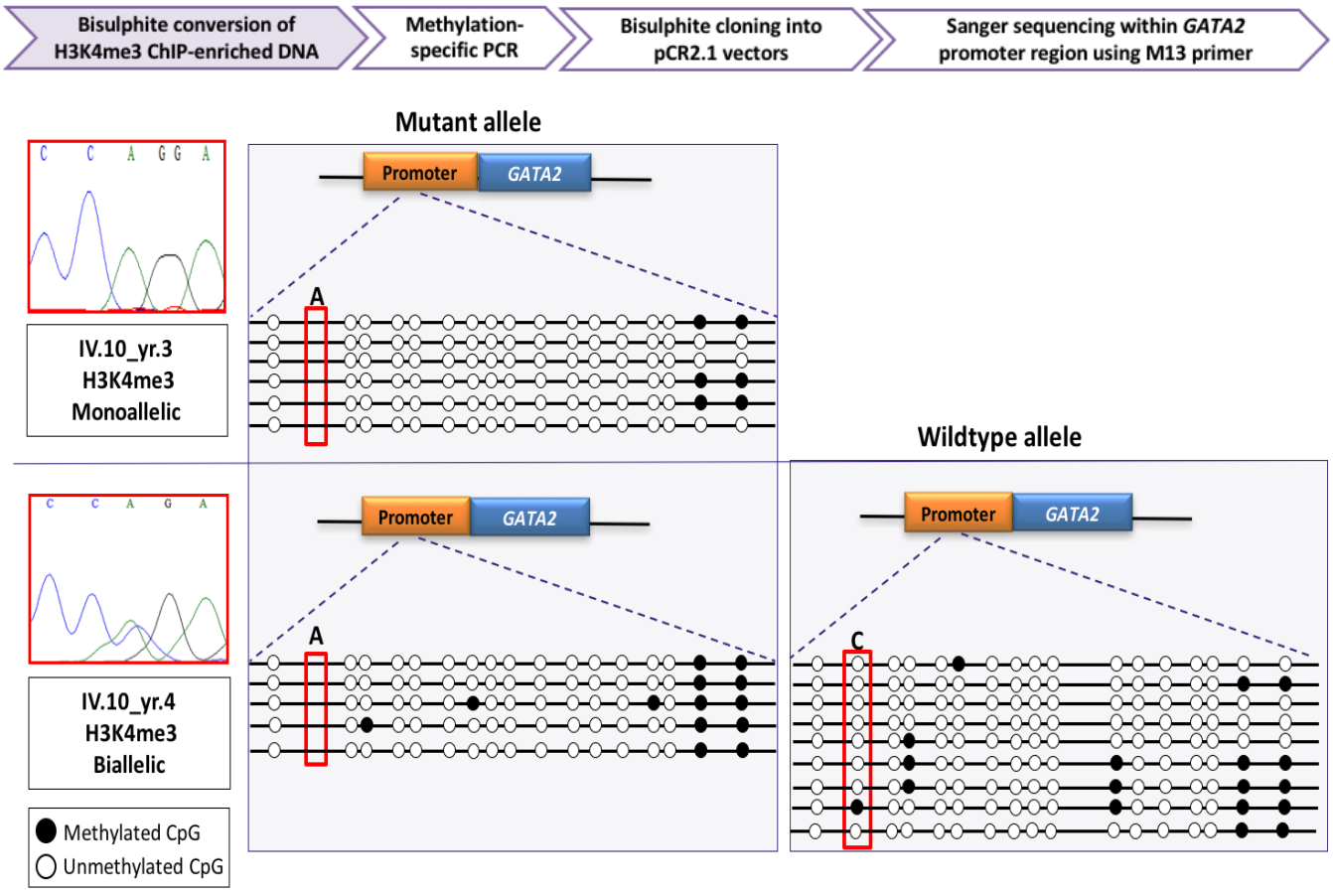


Figure 4.16 Linking DNA methylation and H3K4me3 promoter deposition. Bisulphite-specific PCR followed by cloning and Sanger sequencing covering *GATA2* second promoter SNP [C/A] region overlapping a CpG island was performed to assess DNA methylation patterns of H3K4me3 ChIP-enriched DNA across 2 time-points of our symptomatic patient IV.10 (yr.3 with monoallelic *GATA2* expression) where H3K4me3 is only enriched on the mutant allele and (yr.4 with biallelic *GATA2* expression) where both alleles are equally enriched for H3K4me3. Each row represents a separate clone. Black circles denote methylated CpGs while white circles denote unmethylated CpGs.

4.7.2 An overlap exists between DNA methylation and H3K27me3 deposition

We also performed the same set of experiments on bisulphite-modified DNA derived from IV.10 (yr.3 and 4) that has been immunoprecipitated with H3K27me3 repressive chromatin mark. Based on the proportion of methylated CpGs shown in **Figure 4.17**, DNA methylation appears to coincide with H3K27me3 promoter deposition in all sampling time-points of IV.10 suggesting that the two events are not incompatible. This observation is consistent with the fact that gene promoters that are enriched with H3K27me3 are more likely to be methylated during differentiation and carcinogenesis and this can be attributed to the recruitment of PRC1 and PRC2 complexes that methylate H3K27 and silence CpG island associated genes (Statham et al., 2012, Rose and Klose, 2014). Although the association between PRC2 and Tet1 could explain why a subset of CpG islands that are occupied by PRC2 are not subject to DNA methylation (Neri et al., 2013).

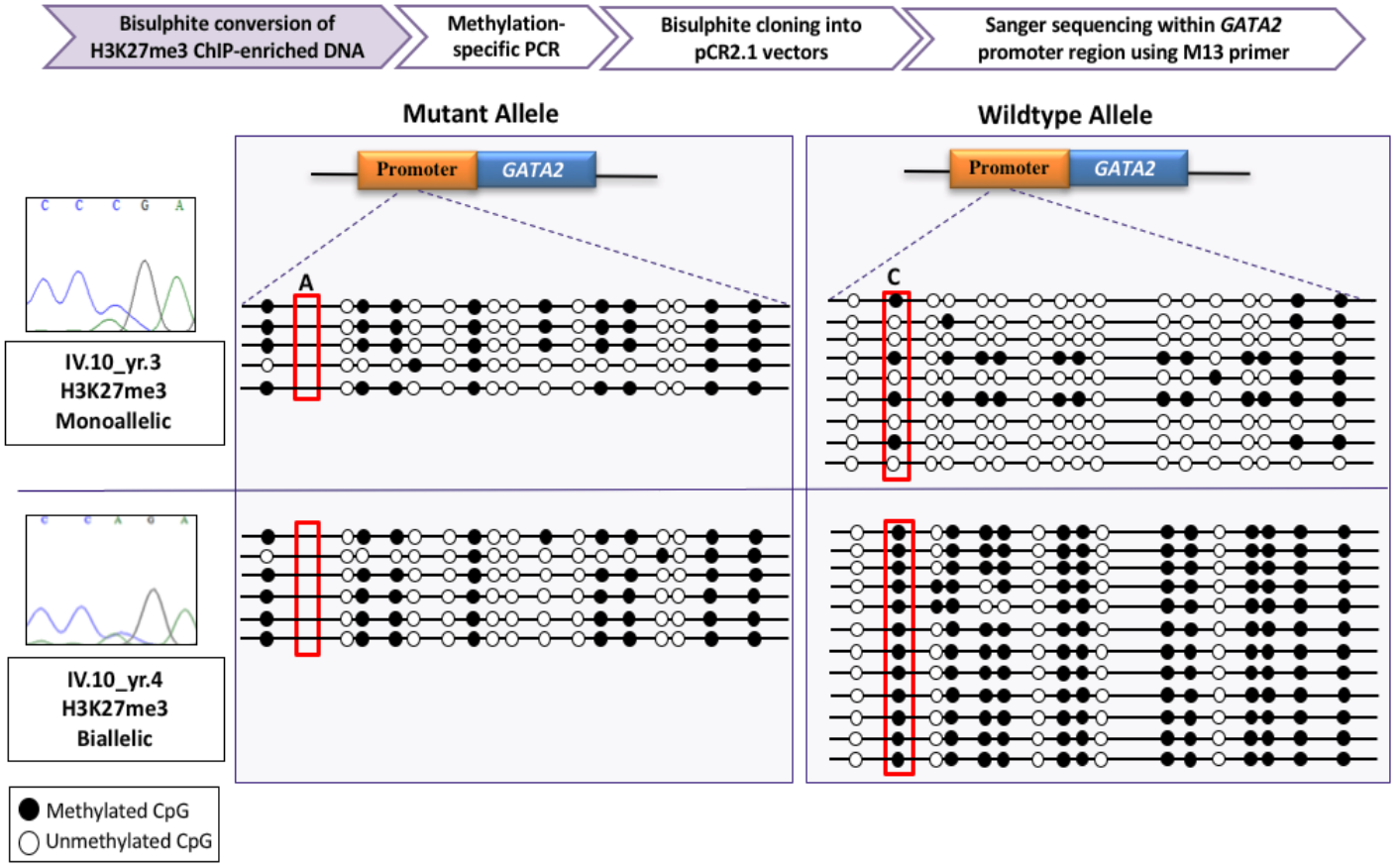


Figure 4.17 Linking DNA methylation and H3K27me3 promoter deposition. Bisulphite-specific PCR followed by cloning and Sanger sequencing covering *GATA2* second promoter SNP [C/A] region overlapping a CpG island was performed to assess DNA methylation patterns of H3K27me3 ChIP-enriched DNA across 2 time-points of our symptomatic patient IV.10 (yr.3 with monoallelic *GATA2* expression) and (yr.4 with biallelic *GATA2* expression). Each row represents a separate clone. Black circles denote methylated CpGs while white circles denote unmethylated CpGs.

4.8 Discussion

The intra- and inter-individual variations in *GATA2* expression profiles seen within and between p.T354M *GATA2*-mutated family members described in **Chapter 3** fostered our investigation focusing on the molecular mechanisms underlying monoallelic *GATA2* expression. We hypothesised that these changes in allelic expression could be mediated by a combination of regulatory and transient epigenetic mechanisms that include changes in DNA methylation and chromatin mark deposition.

The reduction in *GATA2* expression due to allele-specific fluctuations in expression implied that putative *cis*-acting regulatory mechanisms might be involved. Indeed, the identification of a heterozygous *GATA2* promoter 2 SNP residing within the 5'UTR in our symptomatic patient vs. asymptomatic carriers constituted a marker to distinguish between mutant and WT alleles and a first step in establishing the mechanisms governing the observed ASE. Intriguingly, this SNP can also alter (generate/remove) CpG methylation sites within *GATA2* promoter region, resembling a previous study showing that a CpG-SNP (rs12041331) reinforces *PEAR1* enhancer activity in platelet formation through allele-specific DNA methylation (Izzi et al., 2016). Here, we also showed allele-specific DNA methylation as a regulatory mechanism contributing to the silencing of the WT *GATA2* allele in our symptomatic patient's earlier disease time-points. One would postulate promoter methylation to have an impact on TF binding affinity; increased promoter methylation would lead to lower TF binding affinity thereby inhibiting patterns of gene expression and vice versa.

We did not validate allele-specific TF binding occupancy based on the PROMO *in silico* prediction tool in our patient samples. However, our findings are in agreement with a previous study by Celton and colleagues who attributed reduced *GATA2* expression levels in normal karyotype sporadic AMLs (NK-AMLs) to promoter hypermethylation and SNPs acting as loss-of-function mutations, highlighting the importance of epigenetic alterations in modulating gene expression (Celton et al., 2014).

Another notable observation was made following allele-specific ChIP revealing an enrichment of H3K4me3 on the *GATA2* promoter mutant allele compared to the WT allele at diagnosis which was reversed at later follow-up, correlating with reactivation of the WT allele expression. This interpretation is in line with a study by Stern and colleagues who showed mutant *TERT* promoter allele to exhibit H3K4me3 in various cancer cell lines whilst the WT allele retained the H3K27me3 mark of gene silencing (Stern et al., 2015). We also validated that H3K4me3 blocks *de novo* DNA methylation by showing that DNA methylation and H3K4me3 promoter deposition are mutually exclusive in our patient samples. From a translational perspective, this altered *GATA2* allelic expression can be reversed therapeutically by exposure to specific epigenetic inhibitors and/or demethylating agents (e.g. 5-azacytidine). Unfortunately, treating our symptomatic patient cells with KDM5 inhibitor, a drug that inhibits the KDM5 family of histone demethylases, stabilising H3K4me3 levels and could have the potential therefore of reactivating the expression of the silenced WT *GATA2* allele, has proven to be challenging due to the short life of our cells in culture and the scarcity of material available.

Moreover, with the high frequency of promotor 2 SNP (MAF 39%), there would be a reasonable chance that this SNP could be utilised similarly in other *GATA2*-mutated families. Indeed, we should not rule out the possibility that this promoter 2 SNP (rs1806462 [C/A]), by creating an

extra CpG methylation site within the WT allele, could play a direct role in its silencing. To test the contribution of this SNP, we analysed the haplotype of 12 p.Thr354Met *GATA2*-mutated individuals (10 affected members and 2 asymptomatic carriers) from the three families published by Hahn et al. (2011) and we observed two individuals (one symptomatic and one asymptomatic carrier) from two different families heterozygous [C/A] for the SNP (rs1806462). In both cases, the SNP reference allele [C] was *in cis* with the *GATA2* mutation, the opposite of what we detected in our family. Moreover, apart from our symptomatic patient (IV.10), no other family members (including the two deceased MDS/AML cousins IV.1 and IV.6) were heterozygous for these *GATA2* promoter SNPs and given the stable improvement in clinical parameters at IV.10 later time-points (yr.4 and 6), we reasoned that while this CpG-SNP plays a role in *GATA2* monoallelic expression, there is no evidence to establish a correlation between the haplotype rs1806462^{A}-GATA2^{T354M} and the progression of disease/symptomatic status. This SNP was therefore used merely as a vehicle to distinguish between mutant and WT alleles in our DNA methylation and ChIP experiments.

Collectively, our findings from this chapter propose that allele-specific expression of *GATA2* mutant allele is driven by dynamic epigenetic reprogramming; increased DNA methylation linked with lower H3K4me3 promoter deposition on the WT allele and vice versa for the mutant allele (**Figure 4.18**), adding another layer of complexity to the (epi)genetic basis of familial MDS/AML and contributing towards the observed reduced penetrance phenotype seen in certain inherited *GATA2*-mutated MDS/AML families.

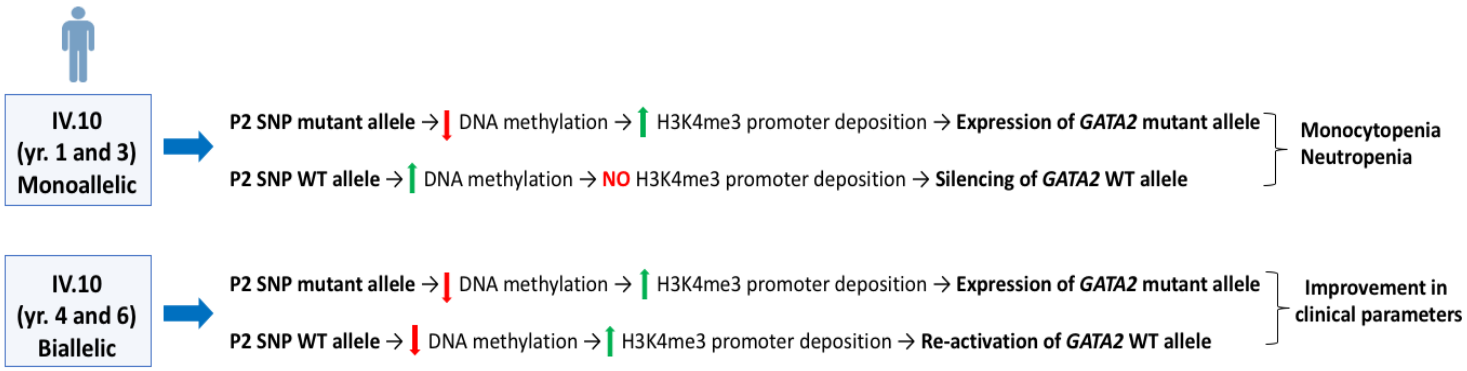


Figure 4.18 Epigenetic dysregulation accounting for the mono-/biallelic *GATA2* expression status observed across the 4 time-points studied in our symptomatic patient (IV.10) and its correlation with disease symptoms, using promoter 2 SNP as a means of distinguishing between mutant and WT alleles.

Chapter 5. Overall Discussion

"Cancer begins and ends with people. In the midst of scientific abstraction, it is sometimes possible to forget this one basic fact…"

– June Goodfield

"The greatest need we have today in the human cancer problem, except for a universal cure, is a method of detecting the presence of cancer before there are any clinical signs of symptoms."

– Sidney Farber

5. Final Discussion

5.1 Introduction

The first example of familial myeloid malignancies was described in the early 1920's but it was not until the discovery of germline *RUNX1* mutations in 6 FPD/AML families in 1999 that our understanding of the mutational landscape and genetic complexity of these inherited syndromes started to develop immensely (Gunz et al., 1975, Song et al., 1999, Akpan et al., 2018). Now two decades later, their recent inclusion as a distinct diagnostic entity in the 2016 WHO classification has renewed efforts to improve the recognition, management and care of this group of patients and their families (Arber et al., 2016). The clinical and molecular heterogeneity of familial MDS/AML, uncovered by harnessing the capabilities of NGS technologies, is reflected in our group's sample collection here at the Barts Cancer Institute in London. Following the discovery of germline mutations in *CEBPA* (Smith et al., 2004), germline mutations have now been identified in ~13 additional loci, including *GATA2,* the subject of this thesis. These studies are not merely important for the respective families themselves but also for sporadic MDS/AML patients as they offer valuable insights into the aetiology of myeloid malignancies and in this thesis, a means to explore the molecular basis of disease penetrance and fluctuations in clinical presentation observed in patients with identical predisposing mutations.

5.2 Same germline mutations, different clinical manifestations – we GATA ask why

We have noted that reduced penetrance is a feature amongst certain *GATA2*-mutated MDS/AML families, especially those carrying germline missense mutations (e.g. p.Thr354Met) and indeed, when we examined all our pedigrees, asymptomatic "silent" mutation carriers are detectable well into their adult years, suggesting that reduced penetrance (long latency; variable expressivity) is an element of these diseases. True estimations of penetrance are however hindered by limited genetic and phenotypic screening of extended family members and incomplete patient/family medical histories, making the prediction of an individual's disease risk and that of future generations particularly challenging. This is further exacerbated by the scarcity of testing tools within diagnostic laboratories worldwide to confirm disease segregation and/or germline mutation status, leading to missed opportunities in detecting and managing these patients and their families, particularly when it comes to appropriate HSCT timing and donor selection.

Herein, we investigated one such family (Bodor et al., 2012) where three generation mutation carriers include both symptomatic and asymptomatic family members and most strikingly, one particular member (IV.10) presented with recurrent infections and significant monocytopenia and neutropenia at the age of 31 years old which fortunately stabilised 3 years after presentation as symptoms improved. We had no satisfactory explanation for the reduced penetrance and the inter- and intra-individual variability in symptoms observed in this family and indeed in the other *GATA2*-mutated MDS/AML families initially reported by (Hahn et al., 2011) and so this thesis has opened up a new research arena not fully addressed in the existing familial leukaemia literature.

We believe this has important connotations for the counselling and management of patients and their families but also potentially widely in relation to predisposition to sporadic disease, where we do not yet have a full grasp of the aetiology and disease initiating events.

5.3 *GATA2* monoallelic expression underlies reduced penetrance in inherited p.T354M-mutated MDS/AML

The results from this thesis showed a novel mechanism to account for the clinical heterogeneity in the setting of germline *GATA2* deficiency. Although it is a single family case study, the careful longitudinal clinical follow-up and temporal molecular analyses set a paradigm for familial leukaemia research and may even have broader implications in MDS/AML as a whole, as it has been challenging to rationalise factors governing mutation acquisition and disease risk; this is now possible considering the in-depth molecular profiling of AML patients using multi-omic approaches such as RNA-seq and the opportunity of detecting monoallelic gene expression. Indeed, this is an ongoing area of investigation in the Fitzgibbon laboratory focusing mainly on sporadic disease, who are employing these technologies to dissect the genetic and biochemical features of poor risk AML (which comprises monosomies, chromosome 3 or 11q abnormalities or complex karyotypes) and capture the personality of a disease that has managed to evade treatment so effectively.

Furthermore, this study highlights the importance of monitoring and evaluating at-risk individuals over time, as access to primary sequential BM or PB material can uncover molecular insights responsible for modulating disease phenotype and particularly in our case example (IV.10), corresponding to changes in *GATA2* expression (total and/or allelic) rather than the more conventional markers reflecting acquisition of somatic mutations e.g. *ASXL1* (**Chapter 3**).

Interestingly, this finding is akin to the skewed *GATA2* ASE observed in sporadic NK-AML patients (Celton et al., 2014) however here we report for the first time its involvement in familial disease. So, whilst we cannot rule out monosomy 7 and *ASXL1* mutations as important secondary events in germline *GATA2* deficiency, as echoed by previous studies (West et al., 2014b, Pastor et al., 2017), the work presented in this thesis suggests that the silencing of the WT *GATA2* allele is perhaps a more critical initiating event and a required step at the very early phases of disease and to drive a patient's initial symptoms. Of note, the mechanisms by which these exact *ASXL1* mutations (p.Gly646TrpfsTer12) and monosomy 7 are selected for in our p.Thr354Met *GATA2*-mutated patients remain unclear and it is not just these aberrations as mutations in *IDH2*, *RUNX1*, *SETBP1* and *STAG2* have also been reported to co-occur with germline *GATA2* mutations in paediatric MDS patients (Ding et al., 2017, Fisher et al., 2017, Wang et al., 2015).

We have a lot to learn and we for example think that the precise nature of the germline mutation itself (e.g. p.Thr354Met) can also impact on disease penetrance and the notion that these missense mutations, by retaining partial protein activity, are insufficient on their own to induce overt malignancy and require the involvement of additional co-operating events to facilitate clonal expansion and proliferation (Kazenwadel et al., 2012, Katsumura et al., 2018) (**Figure 5.1**). This is exemplified by a study reporting the acquisition of somatic *CDC25C*, *JAK2* and *SH2B3* variants in germline *RUNX1*-mutated families (Yoshimi et al., 2014, Tawana et al., 2017b). Nevertheless, whilst mutant ASE of *GATA2* leading to haploinsufficiency is a key initiating event in this family, its value as a predictive biomarker may not necessarily hold true in every familial case but at least it provides a clue going forward that must be considered when assessing families with symptomatic and asymptomatic mutation carriers.

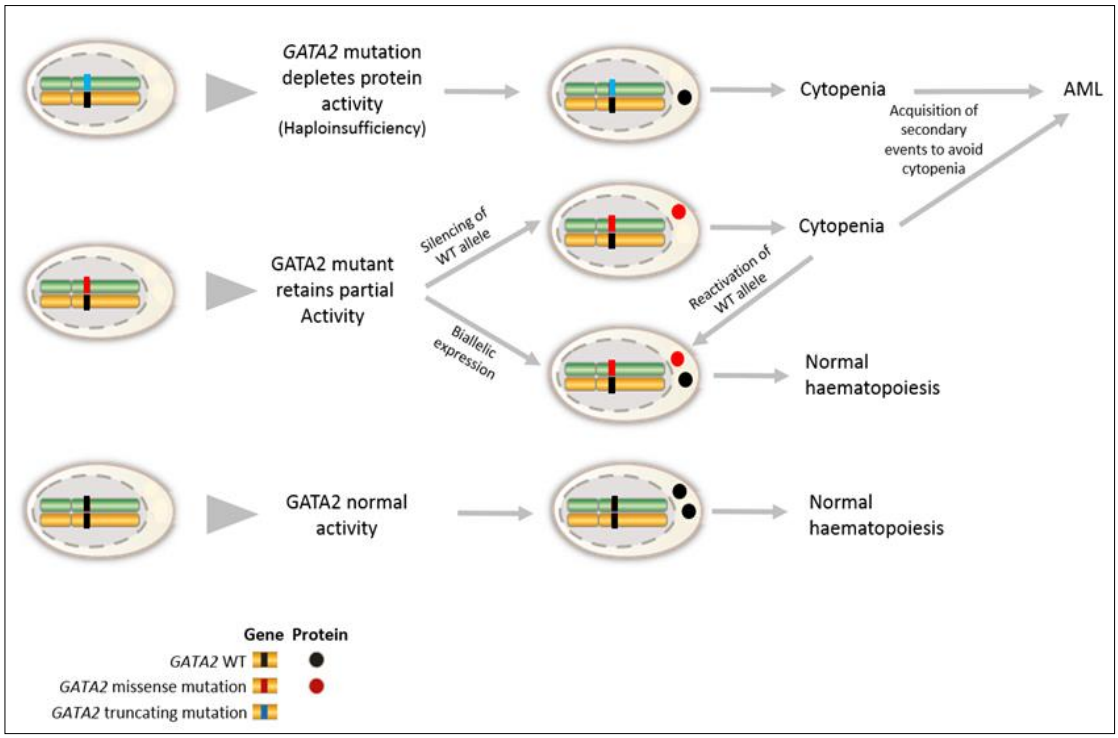


Figure 5.1 Hypothetical model showing the properties of germline *GATA2* mutations (truncating vs. missense vs. WT) and their impact on haematopoietic proliferation and disease state. *GATA2* truncating (e.g. frameshift) mutations (top panel) diminish protein activity, leading to haploinsufficiency, cytopenia and overt disease. In contrast, heterozygous missense *GATA2* mutation (middle panel) can retain partial or residual protein activity, either in relation with DNA binding or interaction with partner proteins, rendering it less disruptive to protein function than *GATA2* truncating or frameshift mutations and therefore may not be enough to drive overt malignancy. The silencing of the WT allele expression, leading to an almost complete depletion of GATA2 function, would therefore be needed to induce the cytopenia and the acquisition of secondary molecular events (e.g. *ASXL1* mutations and monosomy 7) is a way of escaping this cytopenia, leading to AML development. The dynamic nature of this ASE and its correlation with disease state however means that the biallelic *GATA2* expression and normal haematopoiesis can be restored as patients' disease symptoms improve (asymptomatic status). Finally, healthy individuals with 2 WT *GATA2* alleles (bottom panel) have normal haematopoiesis and an uncompromised GATA2 activity.

5.4 Epigenetic alterations regulate monoallelic *GATA2* expression

The second part of the thesis focused on exploring the molecular mechanisms accounting for the allele-specific variations in *GATA2* expression; increased promoter methylation at CpG sites on the WT allele in the earlier disease time-points of our symptomatic patient was noted, along with an allele-specific enrichment of H3K4me3 deposition on the promoter mutant allele, highlighting dynamic epigenetic reprogramming that correlated with disease symptoms (**Chapter 4**). Similarly, a recent study showed that monozygotic twins share the same *de novo GATA2* mutation (c.1192 C>T) but display different phenotypes (the proband was diagnosed with MDS while her twin sister has dysmegakaryopoietic features in the BM), possibly due to differences in their *GATA2* promoter methylation profiles (Kim et al., 2019). Another study on monozygotic twins discordant for childhood leukaemia revealed that they too displayed differential *BRCA1* promoter methylation status (Galetzka et al., 2012), highlighting the role that DNA methylation plays in influencing disease penetrance.

Notably, we showed that these two events (DNA methylation and H3K4me3 promoter deposition) are mutually exclusive within our patient BM cells however it would be challenging to infer how these epigenetic modifications change over time given that the cells were not purified or examined at a single cell level. Even considering this caveat and the transient nature of epigenetic mechanisms, this study highlights the therapeutic potential of modulating *GATA2* expression, either via the use of demethylating agents (e.g. 5-azacytidine) to restore the expression of the silenced allele as described by (Celton et al., 2014) or by maintaining H3K4me3 levels via KDM5 inhibition which warrants further biological studies.

Moreover, we should not underestimate the impact of non-coding regulatory SNPs on the penetrance of coding mutations; this is a widespread genetic phenomenon as evidenced by the wealth of GWAS, with studies demonstrating that high expressing SNP alleles *in cis* can act in concert with the coding mutation to modify disease risk (Lappalainen et al., 2011, Castel et al., 2018). Despite these assertions, here we used a promoter CpG-SNP located within the transcribed 5'UTR region of *GATA2* and overlapping a CpG island as a means of discriminating between mutant and WT alleles, which enabled us to define a haplotype between the promoter SNP allele (**A**) and p.T354M and provided an impetus for further exploration to explain monoallelic *GATA2* expression. However, it is important to state that we cannot exclude the possibility that other SNPs distributed throughout the *GATA2* locus are equally regulating the gene in question.

5.5 Summary

Overall, while we do not know the exact trigger causing *GATA2* ASE in the earlier disease time-points of our symptomatic patient, these findings allowed us to propose a model whereby a pre-MDS stage follows a non-linear trajectory and is likely governed by a complex network of factors regulating gene expression and the acquisition of cytogenetic and/or molecular genetic anomalies over a protracted period of time, predetermined by inherited/host genetic factors (McReynolds et al., 2019). Validating the order of these molecular events experimentally however requires the use of colony forming unit (CFU) assay on $CD34^{+}$ enriched cells which was unattainable in our patient samples due to the low number of blasts. Nevertheless, it is hoped that elucidating these genotype-phenotype associations would improve our understanding of leukaemogeneisis and ultimately provide important insights with which to counsel at-risk mutation carriers and their families.

In summary, it is worth considering that germline mutations are not inherited equally and that variations in expression of these mutated alleles may explain differences in disease outcomes and with this, highlight new opportunities for therapeutic intervention. Altogether, the results presented in this thesis provide a step forward in understanding the molecular mechanisms underpinning reduced penetrance of germline *GATA2* mutations in familial MDS/AML, identifying non-coding regulatory variations and deregulated chromatin signatures alongside additional co-operating somatic mutations and mutant allele-specific expression.

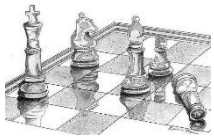
5.6 Potential clinical implications

From a clinical perspective, this study reinforces the importance of careful clinical evaluation of these families and recognition of symptomatic and asymptomatic mutation carriers for close monitoring, genetic counselling, screening and exclusion as potential related HSC transplant donors, since several reports on donor-derived AML act as an important reminder of the potential consequences when these germline mutations are not tested at first hand (Berger et al., 2017, Galera et al., 2018). This study should also urge the haematology community to pause when considering the timing and use of prophylactic allogeneic HSCT in individuals with germline mutations at risk of developing full-blown malignancy, as in the case of IV.10 who has significant leukopenia or in III.5 and III.7 asymptomatic carriers who remain unaffected with no evidence of haematological abnormality (Cuellar-Rodriguez et al., 2011, Akpan et al., 2018). Given the wide spectrum of clinical presentations associated with *GATA2* deficiency syndromes (Hirabayashi et al., 2017), it is pertinent to identify patients early in their disease course when HSCT might be of clinical benefit. Therefore, prognostic biomarkers which can help detect early signs of disease symptoms (e.g. by performing serial gene expression analyses of BM biopsies at presentation) might prove particularly useful and if sufficiently robust, could be translated feasibly into clinical practice.

5.7 Future research directions

"Daring ideas are like chessmen moved forward: they may be beaten, but they may start a winning game."

– Johann Wolfgang von Goethe



There is certainly scope to develop and expand on the information arising from this thesis. In familial disease particularly, there is still a limited understanding of the secondary genetic events that are promoting the occurrence of overt disease and so there is a need to better delineate and decipher the clonal evolutionary dynamics, patterns of mutational co-occurrence and the precise order of genetic monstrosities, whereby potentially a shift in ASE might inform the nature of these events. Consolidating these findings necessitates their replication in additional families and larger patient cohorts. Also, the identification of novel germline predisposing variants with functional and/or translational impact using NGS technologies represents another research avenue that is currently being explored in our laboratory, which relies on sample availability from multiple affected family members with well-annotated clinical information and unknown genetic aetiology (Rio-Machin et al., 2018a). Indeed, while the prevalence of familial MDS/AML at present amounts to ~5% of adults and 4-13% of pediatric patients (Akpan et al., 2018), these percentages are likely to increase as more familial cases and susceptibility genes are being discovered.

Moving forward, we believe that studying familial and sporadic disease in unison represents an important next step in the field as a whole. This is particularly pertinent given the development of mutational risk scores for AML and MDS such as (Gerstung et al., 2017, Nazha et al., 2017) that do not factor in the possibility of familial disease. Nevertheless, considering the increasing vigilance on behalf of the wider haematology community to recognise inherited forms of these

blood cancers, it seems sensible that targeted myeloid resequencing panels in the future include tumour and germline/remission testing for both sporadic and familial disease simultaneously and should be offered in routine diagnostics to all MDS/AML patients, irrespective of their family history. Indeed, we are beginning to learn that familial studies can highlight novel lesions unique to inherited MDS/AML (e.g. *DDX41*), shedding light on new disease biology (Polprasert et al., 2015).

The rarity of these familial cases however means that new gene discovery is likely to require the collaborative effort of an international clinical and research consortia, to power and validate these studies. Ongoing work includes functional analyses of candidate genes using *in vitro* and/or *in vivo* disease model systems (e.g. zebrafish) and gene knockdown techniques (e.g. CRISPR-Cas9) to assess the functional impact of these novel lesions in the pathophysiology of MDS/AML. The possibility exists that germline mutations within the non-coding genome may also be an important component in the aetiology of these diseases, as demonstrated by mutations in the intronic *GATA2* enhancer element (Hsu et al., 2013). Likewise, and based on the findings from this thesis, epigenetic alterations of non-coding regions need to be explored further, with the possibility that gene silencing by means of promoter hypermethylation, as described in *CEBPA* (Fasan et al., 2013a), *GATA2* (Celton et al., 2014) and *RUNX1* (Webber et al., 2013), may be more widely implicated in familial and sporadic MDS/AML.

***

References

ABDEL-WAHAB, O., ADLI, M., LAFAVE, L. M., GAO, J., HRICIK, T., SHIH, A. H., PANDEY, S., PATEL, J. P., CHUNG, Y. R., KOCHE, R., PERNA, F., ZHAO, X., TAYLOR, J. E., PARK, C. Y., CARROLL, M., MELNICK, A., NIMER, S. D., JAFFE, J. D., AIFANTIS, I., BERNSTEIN, B. E. & LEVINE, R. L. 2012. ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer Cell,* 22**,** 180-93.

AKPAN, I. J., OSMAN, A. E. G., DRAZER, M. W. & GODLEY, L. A. 2018. Hereditary Myelodysplastic Syndrome and Acute Myeloid Leukemia: Diagnosis, Questions, and Controversies. *Curr Hematol Malig Rep*.

AL SERAIHI, A. F., RIO-MACHIN, A., TAWANA, K., BÖDÖR, C., WANG, J., NAGANO, A., HEWARD, J. A., IQBAL, S., BEST, S., LEA, N., MCLORNAN, D., KOZYRA, E. J., WLODARSKI, M. W., NIEMEYER, C. M., SCOTT, H., HAHN, C., ELLISON, A., TUMMALA, H., CARDOSO, S. R., VULLIAMY, T., DOKAL, I., BUTLER, T., SMITH, M., CAVENAGH, J. & FITZGIBBON, J. 2018. GATA2 monoallelic expression underlies reduced penetrance in inherited GATA2-mutated MDS/AML. *Leukemia*.

ANDERS, S., PYL, P. T. & HUBER, W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics,* 31**,** 166-9.

ARBER, D. A., ORAZI, A., HASSERJIAN, R., THIELE, J., BOROWITZ, M. J., LE BEAU, M. M., BLOOMFIELD, C. D., CAZZOLA, M. & VARDIMAN, J. W. 2016. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood,* 127**,** 2391-405.

ASADA, S. & KITAMURA, T. 2018. Aberrant histone modifications induced by mutant ASXL1 in myeloid neoplasms. *Int J Hematol*.

BABUSHOK, D. V., BESSLER, M. & OLSON, T. S. 2016. Genetic predisposition to myelodysplastic syndrome and acute myeloid leukemia in children and young adults. *Leuk Lymphoma,* 57**,** 520-36.

BADANO, J. L. & KATSANIS, N. 2002. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet,* 3**,** 779-89.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BANNON, S. A. & DINARDO, C. D. 2016. Hereditary Predispositions to Myelodysplastic Syndrome. *Int J Mol Sci,* 17.

BARBER, R. D., HARMER, D. W., COLEMAN, R. A. & CLARK, B. J. 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics,* 21**,** 389-95.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BAUDHUIN, L. M., FUNKE, B. H., BEAN, L. H., DEIGNAN, J. L., HOFHERR, S., MILLER, D. T., NAGAN, N., SANTANI, A. & SAUNDERS, C. 2016. Classifying Germline Sequence Variants in the Era of Next-Generation Sequencing. *Clin Chem,* 62**,** 799-806.

BECKER, P. 2016. Clonal Hematopoiesis: The Seeds of Leukemia or Innocuous Bystander? *The Hematologist,* 13**,** 7.

BEJAR, R. & STEENSMA, D. P. 2014. Recent developments in myelodysplastic syndromes. *Blood,* 124**,** 2793-803.

BEJAR, R., STEVENSON, K., ABDEL-WAHAB, O., GALILI, N., NILSSON, B., GARCIA-MANERO, G., KANTARJIAN, H., RAZA, A., LEVINE, R. L., NEUBERG, D. & EBERT, B. L. 2011. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med,* 364**,** 2496-506.

BERGER, G., VAN DEN BERG, E., SIKKEMA-RADDATZ, B., ABBOTT, K. M., SINKE, R. J., BUNGENER, L. B., MULDER, A. B. & VELLENGA, E. 2017. Re-emergence of acute myeloid leukemia in donor cells following allogeneic transplantation in a family with a germline DDX41 mutation. *Leukemia,* 31**,** 520-522.

BERGER, S. L., KOUZARIDES, T., SHIEKHATTAR, R. & SHILATIFARD, A. 2009. An operational definition of epigenetics. *Genes Dev,* 23**,** 781-3.

BERI-DEXHEIMER, M., LATGER-CANNARD, V., PHILIPPE, C., BONNET, C., CHAMBON, P., ROTH, V., GREGOIRE, M. J., BORDIGONI, P., LECOMPTE, T., LEHEUP, B. & JONVEAUX, P. 2008. Clinical phenotype of germline RUNX1 haploinsufficiency: from point mutations to large genomic deletions. *Eur J Hum Genet,* 16**,** 1014-8.

BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R., WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125**,** 315-26.

BHALLA, K. N. 2005. Epigenetic and chromatin modifiers as targeted therapy of hematologic malignancies. *J Clin Oncol,* 23**,** 3971-93.

BLACK, J. C., VAN RECHEM, C. & WHETSTINE, J. R. 2012. Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Mol Cell,* 48**,** 491-507.

BLUTEAU, O., SEBERT, M., LEBLANC, T., PEFFAULT DE LATOUR, R., QUENTIN, S., LAINEY, E., HERNANDEZ, L., DALLE, J. H., SICRE DE FONTBRUNE, F., LENGLINE, E., ITZYKSON, R., CLAPPIER, E., BOISSEL, N., VASQUEZ, N., DA COSTA, M., MASLIAH-PLANCHON, J., CUCCUINI, W., RAIMBAULT, A., DE JAEGERE, L., ADES, L., FENAUX, P., MAURY, S., SCHMITT, C., MULLER, M., DOMENECH, C., BLIN, N., BRUNO, B., PELLIER, I., HUNAULT, M., BLANCHE, S., PETIT, A., LEVERGER, G., MICHEL, G., BERTRAND, Y., BARUCHEL, A., SOCIE, G. & SOULIER, J. 2017. A landscape of germline mutations in a cohort of inherited bone marrow failure patients. *Blood*.

BODOR, C., RENNEVILLE, A., SMITH, M., CHARAZAC, A., IQBAL, S., ETANCELIN, P., CAVENAGH, J., BARNETT, M. J., KRAMARZOVA, K., KRISHNAN, B., MATOLCSY, A., PREUDHOMME, C., FITZGIBBON, J. & OWEN, C. 2012. Germ-line GATA2 p.THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival. *Haematologica,* 97**,** 890-4.

BOULTWOOD, J., PERRY, J., PELLAGATTI, A., FERNANDEZ-MERCADO, M., FERNANDEZ-SANTAMARIA, C., CALASANZ, M. J., LARRAYOZ, M. J., GARCIA-DELGADO, M., GIAGOUNIDIS, A., MALCOVATI, L., DELLA PORTA, M. G., JADERSTEN, M., KILLICK, S., HELLSTROM-LINDBERG, E., CAZZOLA, M. & WAINSCOAT, J. S. 2010. Frequent mutation of the polycomb-associated gene ASXL1 in the myelodysplastic syndromes and in acute myeloid leukemia. *Leukemia,* 24**,** 1062-5.

BREEMS, D. A., VAN PUTTEN, W. L., HUIJGENS, P. C., OSSENKOPPELE, G. J., VERHOEF, G. E., VERDONCK, L. F., VELLENGA, E., DE GREEF, G. E., JACKY, E., VAN DER LELIE, J., BOOGAERTS, M. A. & LOWENBERG, B. 2005. Prognostic index for adult patients with acute myeloid leukemia in first relapse. *J Clin Oncol,* 23**,** 1969-78.

BROWN, A. L., CHURPEK, J. E., MALCOVATI, L., DOHNER, H. & GODLEY, L. A. 2017. Recognition of familial myeloid neoplasia in adults. *Semin Hematol,* 54**,** 60-68.

BUCHNER, T., BERDEL, W. E., HAFERLACH, C., HAFERLACH, T., SCHNITTGER, S., MULLER-TIDOW, C., BRAESS, J., SPIEKERMANN, K., KIENAST, J., STAIB, P., GRUNEISEN, A., KERN, W., REICHLE, A., MASCHMEYER, G., AUL, C., LENGFELDER, E., SAUERLAND, M. C., HEINECKE, A., WORMANN, B. & HIDDEMANN, W. 2009. Age-related risk profile and chemotherapy dose response in acute myeloid leukemia: a study by the German Acute Myeloid Leukemia Cooperative Group. *J Clin Oncol,* 27**,** 61-9.

BULLINGER, L., DOHNER, K. & DOHNER, H. 2017. Genomics of Acute Myeloid Leukemia Diagnosis and Pathways. *J Clin Oncol,* 35**,** 934-946.

CANCER GENOME ATLAS RESEARCH, N. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med,* 368**,** 2059-74.

CARDOSO, S. R., ELLISON, A. C. M., WALNE, A. J., CASSIMAN, D., RAGHAVAN, M., KISHORE, B., ANCLIFF, P., RODRIGUEZ-VIGIL, C., DOBBELS, B., RIO-MACHIN, A., AL SERAIHI, A. F. H., PONTIKOS, N., TUMMALA, H., VULLIAMY, T. & DOKAL, I. 2017. Myelodysplasia and liver disease extend the spectrum of RTEL1 related telomeropathies. *Haematologica,* 102**,** e293-e296.

CARDOSO, S. R., RYAN, G., WALNE, A. J., ELLISON, A., LOWE, R., TUMMALA, H., RIO-MACHIN, A., COLLOPY, L., AL SERAIHI, A., WALLIS, Y., PAGE, P., AKIKI, S., FITZGIBBON, J., VULLIAMY, T. & DOKAL, I. 2016. Germline heterozygous DDX41 variants in a subset of familial myelodysplasia and acute myeloid leukemia. *Leukemia*.

CASTAIGNE, S., CHOMIENNE, C., DANIEL, M. T., BALLERINI, P., BERGER, R., FENAUX, P. & DEGOS, L. 1990. All-trans retinoic acid as a differentiation therapy for acute promyelocytic leukemia. I. Clinical results. *Blood,* 76**,** 1704-9.

CASTEL, S. E., CERVERA, A., MOHAMMADI, P., AGUET, F., REVERTER, F., WOLMAN, A., GUIGO, R., IOSSIFOV, I., VASILEVA, A. & LAPPALAINEN, T. 2018. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet,* 50**,** 1327-1334.

CAZZOLA, M., DELLA PORTA, M. G. & MALCOVATI, L. 2013a. The genetic basis of myelodysplasia and its clinical relevance. *Blood,* 122**,** 4021-34.

CAZZOLA, M., ROSSI, M., MALCOVATI, L. & ASSOCIAZIONE ITALIANA PER LA RICERCA SUL CANCRO GRUPPO ITALIANO MALATTIE, M. 2013b. Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms. *Blood,* 121**,** 260-9.

CELTON, M., FOREST, A., GOSSE, G., LEMIEUX, S., HEBERT, J., SAUVAGEAU, G. & WILHELM, B. T. 2014. Epigenetic regulation of GATA2 and its impact on normal karyotype acute myeloid leukemia. *Leukemia,* 28**,** 1617-26.

CHEAH, J. J. C., HAHN, C. N., HIWASE, D. K., SCOTT, H. S. & BROWN, A. L. 2017. Myeloid neoplasms with germline DDX41 mutation. *Int J Hematol,* 106**,** 163-174.

CHEN, D. H., BELOW, J. E., SHIMAMURA, A., KEEL, S. B., MATSUSHITA, M., WOLFF, J., SUL, Y., BONKOWSKI, E., CASTELLA, M., TANIGUCHI, T., NICKERSON, D., PAPAYANNOPOULOU, T., BIRD, T. D. & RASKIND, W. H. 2016. Ataxia-Pancytopenia Syndrome Is Caused by Missense Mutations in SAMD9L. *Am J Hum Genet,* 98**,** 1146-1158.

CHEN, S. & PARMIGIANI, G. 2007. Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol,* 25**,** 1329-33.

CHONG, C. E., VENUGOPAL, P., STOKES, P. H., LEE, Y. K., BRAUTIGAN, P. J., YEUNG, D. T. O., BABIC, M., ENGLER, G. A., LANE, S. W., KLINGLER-HOFFMANN, M., MATTHEWS, J. M., D'ANDREA, R. J., BROWN, A. L., HAHN, C. N. & SCOTT, H. S. 2017. Differential effects on gene transcription and hematopoietic differentiation correlate with GATA2 mutant disease phenotypes. *Leukemia*.

CHURPEK, J. E., LORENZ, R., NEDUMGOTTIL, S., ONEL, K., OLOPADE, O. I., SORRELL, A., OWEN, C. J., BERTUCH, A. A. & GODLEY, L. A. 2013. Proposal for the clinical detection and management of patients and their family members with familial myelodysplastic syndrome/acute leukemia predisposition syndromes. *Leuk Lymphoma,* 54**,** 28-35.

COLLIN, M., DICKINSON, R. & BIGLEY, V. 2015. Haematopoietic and immune defects associated with GATA2 mutation. *Br J Haematol,* 169**,** 173-87.

COOPER, D. N., KRAWCZAK, M., POLYCHRONAKOS, C., TYLER-SMITH, C. & KEHRER-SAWATZKI, H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet,* 132**,** 1077-130.

CORTELLINO, S., XU, J., SANNAI, M., MOORE, R., CARETTI, E., CIGLIANO, A., LE COZ, M., DEVARAJAN, K., WESSELS, A., SOPRANO, D., ABRAMOWITZ, L. K., BARTOLOMEI, M. S., RAMBOW, F., BASSI, M. R., BRUNO, T., FANCIULLI, M., RENNER, C., KLEIN-SZANTO, A. J., MATSUMOTO, Y., KOBI, D., DAVIDSON, I., ALBERTI,

C., LARUE, L. & BELLACOSA, A. 2011. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell,* 146**,** 67-79.

CORTES-LAVAUD, X., LANDECHO, M. F., MAICAS, M., URQUIZA, L., MERINO, J., MORENO-MIRALLES, I. & ODERO, M. D. 2015. GATA2 germline mutations impair GATA2 transcription, causing haploinsufficiency: functional analysis of the p.Arg396Gln mutation. *J Immunol,* 194**,** 2190-8.

COUCH, F. J., WANG, X., MCGUFFOG, L., LEE, A., OLSWOLD, C., KUCHENBAECKER, K. B., SOUCY, P., FREDERICKSEN, Z., BARROWDALE, D., DENNIS, J., GAUDET, M. M., DICKS, E., KOSEL, M., HEALEY, S., SINILNIKOVA, O. M., LEE, A., BACOT, F., VINCENT, D., HOGERVORST, F. B., PEOCK, S., STOPPA-LYONNET, D., JAKUBOWSKA, A., KCONFAB, I., RADICE, P., SCHMUTZLER, R. K., SWE, B., DOMCHEK, S. M., PIEDMONTE, M., SINGER, C. F., FRIEDMAN, E., THOMASSEN, M., ONTARIO CANCER GENETICS, N., HANSEN, T. V., NEUHAUSEN, S. L., SZABO, C. I., BLANCO, I., GREENE, M. H., KARLAN, B. Y., GARBER, J., PHELAN, C. M., WEITZEL, J. N., MONTAGNA, M., OLAH, E., ANDRULIS, I. L., GODWIN, A. K., YANNOUKAKOS, D., GOLDGAR, D. E., CALDES, T., NEVANLINNA, H., OSORIO, A., TERRY, M. B., DALY, M. B., VAN RENSBURG, E. J., HAMANN, U., RAMUS, S. J., TOLAND, A. E., CALIGO, M. A., OLOPADE, O. I., TUNG, N., CLAES, K., BEATTIE, M. S., SOUTHEY, M. C., IMYANITOV, E. N., TISCHKOWITZ, M., JANAVICIUS, R., JOHN, E. M., KWONG, A., DIEZ, O., BALMANA, J., BARKARDOTTIR, R. B., ARUN, B. K., RENNERT, G., TEO, S. H., GANZ, P. A., CAMPBELL, I., VAN DER HOUT, A. H., VAN DEURZEN, C. H., SEYNAEVE, C., GOMEZ GARCIA, E. B., VAN LEEUWEN, F. E., MEIJERS-HEIJBOER, H. E., GILLE, J. J., AUSEMS, M. G., BLOK, M. J., LIGTENBERG, M. J., ROOKUS, M. A., DEVILEE, P., VERHOEF, S., VAN OS, T. A., WIJNEN, J. T., HEBON, EMBRACE, FROST, D., ELLIS, S., FINEBERG, E., PLATTE, R., EVANS, D. G., IZATT, L., EELES, R. A., ADLARD, J., et al. 2013. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet,* 9**,** e1003212.

CRUIJSEN, M., LUBBERT, M., WIJERMANS, P. & HULS, G. 2014. Clinical Results of Hypomethylating Agents in AML Treatment. *J Clin Med,* 4**,** 1-17.

CUELLAR-RODRIGUEZ, J., GEA-BANACLOCHE, J., FREEMAN, A. F., HSU, A. P., ZERBE, C. S., CALVO, K. R., WILDER, J., KURLANDER, R., OLIVIER, K. N., HOLLAND, S. M. & HICKSTEIN, D. D. 2011. Successful allogeneic hematopoietic stem cell transplantation for GATA2 deficiency. *Blood,* 118**,** 3715-20.

DAIGLE, S. R., OLHAVA, E. J., THERKELSEN, C. A., BASAVAPATHRUNI, A., JIN, L., BORIACK-SJODIN, P. A., ALLAIN, C. J., KLAUS, C. R., RAIMONDI, A., SCOTT, M. P., WATERS, N. J., CHESWORTH, R., MOYER, M. P., COPELAND, R. A., RICHON, V. M. & POLLOCK, R. M. 2013. Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. *Blood,* 122**,** 1017-25.

DAVIDSSON, J., PUSCHMANN, A., TEDGARD, U., BRYDER, D., NILSSON, L. & CAMMENGA, J. 2018. SAMD9 and SAMD9L in inherited predisposition to ataxia, pancytopenia, and myeloid malignancies. *Leukemia,* 32**,** 1106-1115.

DAWSON, M. A. & KOUZARIDES, T. 2012. Cancer epigenetics: from mechanism to therapy. *Cell,* 150**,** 12-27.

DELHOMMEAU, F., DUPONT, S., DELLA VALLE, V., JAMES, C., TRANNOY, S., MASSE, A., KOSMIDER, O., LE COUEDIC, J. P., ROBERT, F., ALBERDI, A., LECLUSE, Y., PLO, I., DREYFUS, F. J., MARZAC, C., CASADEVALL, N., LACOMBE, C., ROMANA, S. P., DESSEN, P., SOULIER, J., VIGUIE, F., FONTENAY, M., VAINCHENKER, W. & BERNARD, O. A. 2009. Mutation in TET2 in myeloid cancers. *N Engl J Med,* 360**,** 2289-301.

DESAI, P., MENCIA-TRINCHANT, N., SAVENKOV, O., SIMON, M. S., CHEANG, G., LEE, S., SAMUEL, M., RITCHIE, E. K., GUZMAN, M. L., BALLMAN, K. V., ROBOZ, G. J. & HASSANE, D. C. 2018. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat Med,* 24**,** 1015-1023.

DICKINSON, R. E., GRIFFIN, H., BIGLEY, V., REYNARD, L. N., HUSSAIN, R., HANIFFA, M., LAKEY, J. H., RAHMAN, T., WANG, X. N., MCGOVERN, N., PAGAN, S., COOKSON, S., MCDONALD, D., CHUA, I., WALLIS, J., CANT, A., WRIGHT, M., KEAVNEY, B., CHINNERY, P. F., LOUGHLIN, J., HAMBLETON, S., SANTIBANEZ-KOREF, M. & COLLIN, M. 2011. Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood,* 118**,** 2656-8.

DINARDO, C. D., ROUTBORT, M. J., BANNON, S. A., BENTON, C. B., TAKAHASHI, K., KORNBLAU, S. M., LUTHRA, R., KANAGAL-SHAMANNA, R., MEDEIROS, L. J., GARCIA-MANERO, G., H, M. K., FUTREAL, P. A., MERIC-BERNSTAM, F. & PATEL, K. P. 2018. Improving the detection of patients with inherited predispositions to hematologic malignancies using next-generation sequencing-based leukemia prognostication panels. *Cancer,* 124**,** 2704-2713.

DING, L., LEY, T. J., LARSON, D. E., MILLER, C. A., KOBOLDT, D. C., WELCH, J. S., RITCHEY, J. K., YOUNG, M. A., LAMPRECHT, T., MCLELLAN, M. D., MCMICHAEL, J. F., WALLIS, J. W., LU, C., SHEN, D., HARRIS, C. C., DOOLING, D. J., FULTON, R. S., FULTON, L. L., CHEN, K., SCHMIDT, H., KALICKI-VEIZER, J., MAGRINI, V. J., COOK, L., MCGRATH, S. D., VICKERY, T. L., WENDL, M. C., HEATH, S., WATSON, M. A., LINK, D. C., TOMASSON, M. H., SHANNON, W. D., PAYTON, J. E., KULKARNI, S., WESTERVELT, P., WALTER, M. J., GRAUBERT, T. A., MARDIS, E. R., WILSON, R. K. & DIPERSIO, J. F. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature,* 481**,** 506-10.

DING, L. W., IKEZOE, T., TAN, K. T., MORI, M., MAYAKONDA, A., CHIEN, W., LIN, D. C., JIANG, Y. Y., LILL, M., YANG, H., SUN, Q. Y. & KOEFFLER, H. P. 2017. Mutational profiling of a MonoMAC syndrome family with GATA2 deficiency. *Leukemia,* 31**,** 244-245.

DOHNER, H., ESTEY, E., GRIMWADE, D., AMADORI, S., APPELBAUM, F. R., BUCHNER, T., DOMBRET, H., EBERT, B. L., FENAUX, P., LARSON, R. A., LEVINE, R. L., LO-COCO, F., NAOE, T., NIEDERWIESER, D., OSSENKOPPELE, G. J., SANZ, M., SIERRA, J., TALLMAN, M. S., TIEN, H. F., WEI, A. H., LOWENBERG, B. & BLOOMFIELD, C. D. 2017. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood,* 129**,** 424-447.

DOHNER, K. & PASCHKA, P. 2014. Intermediate-risk acute myeloid leukemia therapy: current and future. *Hematology Am Soc Hematol Educ Program,* 2014**,** 34-43.

DOKAL, I. & VULLIAMY, T. 2010. Inherited bone marrow failure syndromes. *Haematologica,* 95**,** 1236-40.

DOMBRET, H. & GARDIN, C. 2016. An update of current treatments for adult acute myeloid leukemia. *Blood,* 127**,** 53-61.

DOMBRET, H., SEYMOUR, J. F., BUTRYM, A., WIERZBOWSKA, A., SELLESLAG, D., JANG, J. H., KUMAR, R., CAVENAGH, J., SCHUH, A. C., CANDONI, A., RECHER, C., SANDHU, I., BERNAL DEL CASTILLO, T., AL-ALI, H. K., MARTINELLI, G., FALANTES, J., NOPPENEY, R., STONE, R. M., MINDEN, M. D., MCINTYRE, H., SONGER, S., LUCY, L. M., BEACH, C. L. & DOHNER, H. 2015. International phase 3 study of azacitidine vs conventional care regimens in older patients with newly diagnosed AML with >30% blasts. *Blood,* 126**,** 291-9.

DRAZER, M., FEURSTEIN, S., WEST, A., JONES, M., CHURPEK, J. & GODLEY, L. 2016. How I diagnose and manage individuals at risk for inherited myeloid malignancies. *Blood,* 128**,** 1800-1813.

DRAZER, M. W., KADRI, S., SUKHANOVA, M., PATIL, S. A., WEST, A. H., FEURSTEIN, S., CALDERON, D. A., JONES, M. F., WEIPERT, C. M., DAUGHERTY, C. K., CEBALLOS-LOPEZ, A. A., RACA, G., LINGEN, M. W., LI, Z., SEGAL, J. P., CHURPEK, J. E. & GODLEY, L. A. 2018. Prognostic tumor sequencing panels frequently identify germ line variants associated with hereditary hematopoietic malignancies. *Blood Adv,* 2**,** 146-150.

DUFOUR, A., SCHNEIDER, F., METZELER, K. H., HOSTER, E., SCHNEIDER, S., ZELLMEIER, E., BENTHAUS, T., SAUERLAND, M. C., BERDEL, W. E., BUCHNER, T., WORMANN, B., BRAESS, J., HIDDEMANN, W., BOHLANDER, S. K. & SPIEKERMANN, K. 2010. Acute myeloid leukemia with biallelic CEBPA gene mutations and normal karyotype represents a distinct genetic entity associated with a favorable clinical outcome. *J Clin Oncol,* 28**,** 570-7.

FASAN, A., ALPERMANN, T., HAFERLACH, C., GROSSMANN, V., ROLLER, A., KOHLMANN, A., EDER, C., KERN, W., HAFERLACH, T. & SCHNITTGER, S. 2013a. Frequency and prognostic impact of CEBPA proximal, distal and core promoter methylation in normal karyotype AML: a study on 623 cases. *PLoS One,* 8**,** e54365.

FASAN, A., EDER, C., HAFERLACH, C., GROSSMANN, V., KOHLMANN, A., DICKER, F., KERN, W., HAFERLACH, T. & SCHNITTGER, S. 2013b. GATA2 mutations are frequent in intermediate-risk karyotype AML with biallelic CEBPA mutations and are associated with favorable prognosis. *Leukemia,* 27**,** 482-5.
FEINBERG, A. P., CUI, H. & OHLSSON, R. 2002. DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. *Semin Cancer Biol,* 12**,** 389-98.

FERRANDO, A. A. & LOPEZ-OTIN, C. 2017. Clonal evolution in leukemia. *Nat Med,* 23**,** 1135-1145.
FICZ, G., BRANCO, M. R., SEISENBERGER, S., SANTOS, F., KRUEGER, F., HORE, T. A., MARQUES, C. J., ANDREWS, S. & REIK, W. 2011. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature,* 473**,** 398-402.

FIGUEROA, M. E., ABDEL-WAHAB, O., LU, C., WARD, P. S., PATEL, J., SHIH, A., LI, Y., BHAGWAT, N., VASANTHAKUMAR, A., FERNANDEZ, H. F., TALLMAN, M. S., SUN, Z., WOLNIAK, K., PEETERS, J. K., LIU, W., CHOE, S. E., FANTIN, V. R., PAIETTA, E., LOWENBERG, B., LICHT, J. D., GODLEY, L. A., DELWEL, R., VALK, P. J., THOMPSON, C. B., LEVINE, R. L. & MELNICK, A. 2010. Leukemic IDH1 and IDH2 mutations result in

a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell,* 18**,** 553-67.

FISHER, K. E., HSU, A. P., WILLIAMS, C. L., SAYEED, H., MERRITT, B. Y., ELGHETANY, M. T., HOLLAND, S. M., BERTUCH, A. A. & GRAMATGES, M. M. 2017. Somatic mutations in children with GATA2-associated myelodysplastic syndrome who lack other features of GATA2 deficiency. *Blood Adv,* 1**,** 443-448.

FORD, A. M., BENNETT, C. A., PRICE, C. M., BRUIN, M. C., VAN WERING, E. R. & GREAVES, M. 1998. Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. *Proc Natl Acad Sci U S A,* 95**,** 4584-8.

FROMMER, M., MCDONALD, L. E., MILLAR, D. S., COLLIS, C. M., WATT, F., GRIGG, G. W., MOLLOY, P. L. & PAUL, C. L. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A,* 89**,** 1827-31.

FUJIWARA, T., FUKUHARA, N., FUNAYAMA, R., NARIAI, N., KAMATA, M., NAGASHIMA, T., KOJIMA, K., ONISHI, Y., SASAHARA, Y., ISHIZAWA, K., NAGASAKI, M., NAKAYAMA, K. & HARIGAE, H. 2014. Identification of acquired mutations by whole-genome sequencing in GATA-2 deficiency evolving into myelodysplasia and acute leukemia. *Ann Hematol,* 93**,** 1515-22.

GADE, P. & KALVAKOLANU, D. V. 2012. Chromatin immunoprecipitation assay as a tool for analyzing transcription factor activity. *Methods Mol Biol,* 809**,** 85-104.

GALE, R. E., GREEN, C., ALLEN, C., MEAD, A. J., BURNETT, A. K., HILLS, R. K., LINCH, D. C. & MEDICAL RESEARCH COUNCIL ADULT LEUKAEMIA WORKING, P. 2008. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood,* 111**,** 2776-84.

GALERA, P., HSU, A. P., WANG, W., DROLL, S., CHEN, R., SCHWARTZ, J. R., KLCO, J. M., ARAI, S., MAESE, L., ZERBE, C., PARTA, M. J., YOUNG, N. S., HOLLAND, S. M., HICKSTEIN, D. D. & CALVO, K. R. 2018. Donor-derived MDS/AML in families with germline GATA2 mutation. *Blood,* 132**,** 1994-1998.

GALETZKA, D., HANSMANN, T., EL HAJJ, N., WEIS, E., IRMSCHER, B., LUDWIG, M., SCHNEIDER-RATZKE, B., KOHLSCHMIDT, N., BEYER, V., BARTSCH, O., ZECHNER, U., SPIX, C. & HAAF, T. 2012. Monozygotic twins discordant for constitutive BRCA1 promoter methylation, childhood cancer and secondary cancer. *Epigenetics,* 7**,** 47-54.

GAO, J., CHEN, Y. H. & PETERSON, L. C. 2015. GATA family transcriptional factors: emerging suspects in hematologic disorders. *Exp Hematol Oncol,* 4**,** 28.

GARCIA, J. S. & STONE, R. M. 2017. The Development of FLT3 Inhibitors in Acute Myeloid Leukemia. *Hematol Oncol Clin North Am,* 31**,** 663-680.

GAUDET, M. M., KIRCHHOFF, T., GREEN, T., VIJAI, J., KORN, J. M., GUIDUCCI, C., SEGRE, A. V., MCGEE, K., MCGUFFOG, L., KARTSONAKI, C., MORRISON, J., HEALEY, S., SINILNIKOVA, O. M., STOPPA-LYONNET, D., MAZOYER, S., GAUTHIER-VILLARS, M., SOBOL, H., LONGY, M., FRENAY, M., COLLABORATORS, G. S.,

HOGERVORST, F. B., ROOKUS, M. A., COLLEE, J. M., HOOGERBRUGGE, N., VAN ROOZENDAAL, K. E., COLLABORATORS, H. S., PIEDMONTE, M., RUBINSTEIN, W., NERENSTONE, S., VAN LE, L., BLANK, S. V., CALDES, T., DE LA HOYA, M., NEVANLINNA, H., AITTOMAKI, K., LAZARO, C., BLANCO, I., ARASON, A., JOHANNSSON, O. T., BARKARDOTTIR, R. B., DEVILEE, P., OLOPADE, O. I., NEUHAUSEN, S. L., WANG, X., FREDERICKSEN, Z. S., PETERLONGO, P., MANOUKIAN, S., BARILE, M., VIEL, A., RADICE, P., PHELAN, C. M., NAROD, S., RENNERT, G., LEJBKOWICZ, F., FLUGELMAN, A., ANDRULIS, I. L., GLENDON, G., OZCELIK, H., OCGN, TOLAND, A. E., MONTAGNA, M., D'ANDREA, E., FRIEDMAN, E., LAITMAN, Y., BORG, A., BEATTIE, M., RAMUS, S. J., DOMCHEK, S. M., NATHANSON, K. L., REBBECK, T., SPURDLE, A. B., CHEN, X., HOLLAND, H., KCONFAB, JOHN, E. M., HOPPER, J. L., BUYS, S. S., DALY, M. B., SOUTHEY, M. C., TERRY, M. B., TUNG, N., OVEREEM HANSEN, T. V., NIELSEN, F. C., GREENE, M. H., MAI, P. L., OSORIO, A., DURAN, M., ANDRES, R., BENITEZ, J., WEITZEL, J. N., GARBER, J., HAMANN, U., EMBRACE, PEOCK, S., COOK, M., OLIVER, C., FROST, D., PLATTE, R., EVANS, D. G., LALLOO, F., et al. 2010. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. *PLoS Genet,* 6**,** e1001183.

GENOVESE, G., KAHLER, A. K., HANDSAKER, R. E., LINDBERG, J., ROSE, S. A., BAKHOUM, S. F., CHAMBERT, K., MICK, E., NEALE, B. M., FROMER, M., PURCELL, S. M., SVANTESSON, O., LANDEN, M., HOGLUND, M., LEHMANN, S., GABRIEL, S. B., MORAN, J. L., LANDER, E. S., SULLIVAN, P. F., SKLAR, P., GRONBERG, H., HULTMAN, C. M. & MCCARROLL, S. A. 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med,* 371**,** 2477-87.

GERSTUNG, M., PAPAEMMANUIL, E., MARTINCORENA, I., BULLINGER, L., GAIDZIK, V. I., PASCHKA, P., HEUSER, M., THOL, F., BOLLI, N., GANLY, P., GANSER, A., MCDERMOTT, U., DOHNER, K., SCHLENK, R. F., DOHNER, H. & CAMPBELL, P. J. 2017. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet,* 49**,** 332-340.

GILLILAND, D. G. 2002. Molecular genetics of human leukemias: new insights into therapy. *Semin Hematol,* 39**,** 6-11.

GODLEY, L. A. 2014. Inherited predisposition to acute myeloid leukemia. *Semin Hematol,* 51**,** 306-21.

GRAUBERT, T. A. & MARDIS, E. R. 2011. Genomics of acute myeloid leukemia. *Cancer J,* 17**,** 487-91.

GREEN, C. L., TAWANA, K., HILLS, R. K., BODOR, C., FITZGIBBON, J., INGLOTT, S., ANCLIFF, P., BURNETT, A. K., LINCH, D. C. & GALE, R. E. 2013. GATA2 mutations in sporadic and familial acute myeloid leukaemia patients with CEBPA mutations. *Br J Haematol,* 161**,** 701-5.

GREENBERG, P. L., STONE, R. M., AL-KALI, A., BARTA, S. K., BEJAR, R., BENNETT, J. M., CARRAWAY, H., DE CASTRO, C. M., DEEG, H. J., DEZERN, A. E., FATHI, A. T., FRANKFURT, O., GAENSLER, K., GARCIA-MANERO, G., GRIFFITHS, E. A., HEAD, D., HORSFALL, R., JOHNSON, R. A., JUCKETT, M., KLIMEK, V. M., KOMROKJI, R., KUJAWSKI, L. A., MANESS, L. J., O'DONNELL, M. R., POLLYEA, D. A., SHAMI, P. J., STEIN, B. L., WALKER, A. R., WESTERVELT, P., ZEIDAN, A., SHEAD, D. A. & SMITH, C. 2017. Myelodysplastic Syndromes, Version 2.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw,* 15**,** 60-87.

GREENBERG, P. L., TUECHLER, H., SCHANZ, J., SANZ, G., GARCIA-MANERO, G., SOLE, F., BENNETT, J. M., BOWEN, D., FENAUX, P., DREYFUS, F., KANTARJIAN, H., KUENDGEN, A., LEVIS, A., MALCOVATI, L., CAZZOLA, M., CERMAK, J., FONATSCH, C., LE BEAU, M. M., SLOVAK, M. L., KRIEGER, O., LUEBBERT, M., MACIEJEWSKI, J., MAGALHAES, S. M., MIYAZAKI, Y., PFEILSTOCKER, M., SEKERES, M., SPERR, W. R., STAUDER, R., TAURO, S., VALENT, P., VALLESPI, T., VAN DE LOOSDRECHT, A. A., GERMING, U. & HAASE, D. 2012. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood,* 120**,** 2454-65.

GREIF, P. A., DUFOUR, A., KONSTANDIN, N. P., KSIENZYK, B., ZELLMEIER, E., TIZAZU, B., STURM, J., BENTHAUS, T., HEROLD, T., YAGHMAIE, M., DORGE, P., HOPFNER, K. P., HAUSER, A., GRAF, A., KREBS, S., BLUM, H., KAKADIA, P. M., SCHNEIDER, S., HOSTER, E., SCHNEIDER, F., STANULLA, M., BRAESS, J., SAUERLAND, M. C., BERDEL, W. E., BUCHNER, T., WOERMANN, B. J., HIDDEMANN, W., SPIEKERMANN, K. & BOHLANDER, S. K. 2012. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood,* 120**,** 395-403.

GRIMWADE, D. 2012. The changing paradigm of prognostic factors in acute myeloid leukaemia. *Best Pract Res Clin Haematol,* 25**,** 419-25.

GRIMWADE, D. & FREEMAN, S. D. 2014. Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"? *Hematology Am Soc Hematol Educ Program,* 2014**,** 222-33.
GRIMWADE, D., HILLS, R. K., MOORMAN, A. V., WALKER, H., CHATTERS, S., GOLDSTONE, A. H., WHEATLEY, K., HARRISON, C. J., BURNETT, A. K. & NATIONAL CANCER RESEARCH INSTITUTE ADULT LEUKAEMIA WORKING, G. 2010. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood,* 116**,** 354-65.

GRIMWADE, D., IVEY, A. & HUNTLY, B. J. 2016. Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. *Blood,* 127**,** 29-41.

GRIMWADE, D., WALKER, H., OLIVER, F., WHEATLEY, K., HARRISON, C., HARRISON, G., REES, J., HANN, I., STEVENS, R., BURNETT, A. & GOLDSTONE, A. 1998. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood,* 92**,** 2322-33.

GROSSMANN, V., SCHNITTGER, S., KOHLMANN, A., EDER, C., ROLLER, A., DICKER, F., SCHMID, C., WENDTNER, C. M., STAIB, P., SERVE, H., KREUZER, K. A., KERN, W., HAFERLACH, T. & HAFERLACH, C. 2012. A novel hierarchical prognostic model of AML solely based on molecular mutations. *Blood,* 120**,** 2963-72.

GUIDUGLI, L., JOHNSON, A. K., ALKORTA-ARANBURU, G., NELAKUDITI, V., ARNDT, K., CHURPEK, J. E., GODLEY, L. A., TOWNSLEY, D., YOUNG, N. S., FITZPATRICK, C., DEL GAUDIO, D., DAS, S. & LI, Z. 2017. Clinical utility of gene panel-based testing for hereditary myelodysplastic syndrome/acute leukemia predisposition syndromes. *Leukemia,* 31**,** 1226-1229.

GUNZ, F. W., GUNZ, J. P., VEALE, A. M., CHAPMAN, C. J. & HOUSTON, I. B. 1975. Familial leukaemia: a study of 909 families. *Scand J Haematol,* 15**,** 117-31.

HAHN, C. N., CHONG, C. E., CARMICHAEL, C. L., WILKINS, E. J., BRAUTIGAN, P. J., LI, X. C., BABIC, M., LIN, M., CARMAGNAC, A., LEE, Y. K., KOK, C. H., GAGLIARDI, L., FRIEND, K. L., EKERT, P. G., BUTCHER, C. M., BROWN, A. L., LEWIS, I. D., TO, L. B., TIMMS, A. E., STOREK, J., MOORE, S., ALTREE, M., ESCHER, R., BARDY, P. G., SUTHERS, G. K., D'ANDREA, R. J., HORWITZ, M. S. & SCOTT, H. S. 2011. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet,* 43**,** 1012-7.

HANSEN, K. D., IRIZARRY, R. A. & WU, Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics,* 13**,** 204-16.

HARIKUMAR, A. & MESHORER, E. 2015. Chromatin remodeling and bivalent histone modifications in embryonic stem cells. *EMBO Rep,* 16**,** 1609-19.

HASHIMSHONY, T., ZHANG, J., KESHET, I., BUSTIN, M. & CEDAR, H. 2003. The role of DNA methylation in setting up chromatin structure during development. *Nat Genet,* 34**,** 187-92.

HE, Y. F., LI, B. Z., LI, Z., LIU, P., WANG, Y., TANG, Q., DING, J., JIA, Y., CHEN, Z., LI, L., SUN, Y., LI, X., DAI, Q., SONG, C. X., ZHANG, K., HE, C. & XU, G. L. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science,* 333**,** 1303-7.

HENIKOFF, S. 2000. Heterochromatin function in complex genomes. *Biochim Biophys Acta,* 1470**,** O1-8.

HIRABAYASHI, S., WLODARSKI, M. W., KOZYRA, E. & NIEMEYER, C. M. 2017. Heterogeneity of GATA2-related myeloid neoplasms. *Int J Hematol,* 106**,** 175-182.

HOCK, H. 2012. A complex Polycomb issue: the two faces of EZH2 in cancer. *Genes Dev,* 26**,** 751-5.

HOLLIDAY, R. 1987. DNA methylation and epigenetic defects in carcinogenesis. *Mutat Res,* 181**,** 215-7.

HOLLIDAY, R. & PUGH, J. E. 1975. DNA modification mechanisms and gene activity during development. *Science,* 187**,** 226-32.

HOLME, H., HOSSAIN, U., KIRWAN, M., WALNE, A., VULLIAMY, T. & DOKAL, I. 2012. Marked genetic heterogeneity in familial myelodysplasia/acute myeloid leukaemia. *Br J Haematol,* 158**,** 242-248.

HORAN, J. T., LOGAN, B. R., AGOVI-JOHNSON, M. A., LAZARUS, H. M., BACIGALUPO, A. A., BALLEN, K. K., BREDESON, C. N., CARABASI, M. H., GUPTA, V., HALE, G. A., KHOURY, H. J., JUCKETT, M. B., LITZOW, M. R., MARTINO, R., MCCARTHY, P. L., SMITH, F. O., RIZZO, J. D. & PASQUINI, M. C. 2011. Reducing the risk for transplantation-related mortality after allogeneic hematopoietic cell transplantation: how much progress has been made? *J Clin Oncol,* 29**,** 805-13.

HORWITZ, M. S. 2014. GATA2 deficiency: flesh and blood. *Blood,* 123**,** 799-800.

HSU, A. P., JOHNSON, K. D., FALCONE, E. L., SANALKUMAR, R., SANCHEZ, L., HICKSTEIN, D. D., CUELLAR-RODRIGUEZ, J., LEMIEUX, J. E., ZERBE, C. S., BRESNICK, E. H. & HOLLAND, S. M. 2013. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood,* 121**,** 3830-7, S1-7.

HYDE, R. K. & LIU, P. P. 2011. GATA2 mutations lead to MDS and AML. *Nat Genet,* 43**,** 926-7.

ILLINGWORTH, R., KERR, A., DESOUSA, D., JORGENSEN, H., ELLIS, P., STALKER, J., JACKSON, D., CLEE, C., PLUMB, R., ROGERS, J., HUMPHRAY, S., COX, T., LANGFORD, C. & BIRD, A. 2008. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol,* 6**,** e22.

INOUE, D., KITAURA, J., MATSUI, H., HOU, H. A., CHOU, W. C., NAGAMACHI, A., KAWABATA, K. C., TOGAMI, K., NAGASE, R., HORIKAWA, S., SAIKA, M., MICOL, J. B., HAYASHI, Y., HARADA, Y., HARADA, H., INABA, T., TIEN, H. F., ABDEL-WAHAB, O. & KITAMURA, T. 2015. SETBP1 mutations drive leukemic transformation in ASXL1-mutated MDS. *Leukemia,* 29**,** 847-57.

INVERNIZZI, R., QUAGLIA, F. & PORTA, M. G. 2015. Importance of classical morphology in the diagnosis of myelodysplastic syndrome. *Mediterr J Hematol Infect Dis,* 7**,** e2015035.

IVEY, A., HILLS, R. K., SIMPSON, M. A., JOVANOVIC, J. V., GILKES, A., GRECH, A., PATEL, Y., BHUDIA, N., FARAH, H., MASON, J., WALL, K., AKIKI, S., GRIFFITHS, M., SOLOMON, E., MCCAUGHAN, F., LINCH, D. C., GALE, R. E., VYAS, P., FREEMAN, S. D., RUSSELL, N., BURNETT, A. K., GRIMWADE, D. & GROUP, U. K. N. C. R. I. A. W. 2016. Assessment of Minimal Residual Disease in Standard-Risk AML. *N Engl J Med,* 374**,** 422-33.

IZZI, B., PISTONI, M., CLUDTS, K., AKKOR, P., LAMBRECHTS, D., VERFAILLIE, C., VERHAMME, P., FRESON, K. & HOYLAERTS, M. F. 2016. Allele-specific DNA methylation reinforces PEAR1 enhancer activity. *Blood,* 128**,** 1003-12.

JAFFE, E., HARRIS, N., STEIN, H. & JW., V. 2001. *Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues,* Lyon, France, WHO press.

JONGMANS, M. C., KUIPER, R. P., CARMICHAEL, C. L., WILKINS, E. J., DORS, N., CARMAGNAC, A., SCHOUTEN-VAN MEETEREN, A. Y., LI, X., STANKOVIC, M., KAMPING, E., BENGTSSON, H., SCHOENMAKERS, E. F., VAN KESSEL, A. G., HOOGERBRUGGE, P. M., HAHN, C. N., BRONS, P. P., SCOTT, H. S. & HOOGERBRUGGE, N. 2010. Novel RUNX1 mutations in familial platelet disorder with enhanced risk for acute myeloid leukemia: clues for improved identification of the FPD/AML syndrome. *Leukemia,* 24**,** 242-6.

KADRI, S., LONG, B. C., MUJACIC, I., ZHEN, C. J., WURST, M. N., SHARMA, S., MCDONALD, N., NIU, N., BENHAMED, S., TUTEJA, J. H., SEIWERT, T. Y., WHITE, K. P., MCNERNEY, M. E., FITZPATRICK, C., WANG, Y. L., FURTADO, L. V. & SEGAL, J. P. 2017. Clinical Validation of a Next-Generation Sequencing Genomic Oncology Panel via Cross-Platform Benchmarking against Established Amplicon Sequencing Assays. *J Mol Diagn,* 19**,** 43-56.

KANAMARU, A., TAKEMOTO, Y., TANIMOTO, M., MURAKAMI, H., ASOU, N., KOBAYASHI, T., KURIYAMA, K., OHMOTO, E., SAKAMAKI, H., TSUBAKI, K. & ET AL. 1995. All-trans retinoic acid for the treatment of newly diagnosed acute promyelocytic leukemia. Japan Adult Leukemia Study Group. *Blood,* 85**,** 1202-6.

KANDIMALLA, R., VAN TILBORG, A. A. & ZWARTHOFF, E. C. 2013. DNA methylation-based biomarkers in bladder cancer. *Nat Rev Urol,* 10**,** 327-35.

KATSUMURA, K. R., BRESNICK, E. H. & GROUP, G. F. M. 2017. The GATA factor revolution in hematology. *Blood,* 129**,** 2092-2102

.

KATSUMURA, K. R., MEHTA, C., HEWITT, K. J., SOUKUP, A. A., FRAGA DE ANDRADE, I., RANHEIM, E. A., JOHNSON, K. D. & BRESNICK, E. H. 2018. Human leukemia mutations corrupt but do not abrogate GATA-2 function. *Proc Natl Acad Sci U S A,* 115**,** E10109-E10118.

KAZENWADEL, J., SECKER, G. A., LIU, Y. J., ROSENFELD, J. A., WILDIN, R. S., CUELLAR-RODRIGUEZ, J., HSU, A. P., DYACK, S., FERNANDEZ, C. V., CHONG, C. E., BABIC, M., BARDY, P. G., SHIMAMURA, A., ZHANG, M. Y., WALSH, T., HOLLAND, S. M., HICKSTEIN, D. D., HORWITZ, M. S., HAHN, C. N., SCOTT, H. S. & HARVEY, N. L. 2012. Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature. *Blood,* 119**,** 1283-91.

KIM, E. S. 2017. Enasidenib: First Global Approval. *Drugs,* 77**,** 1705-1711.

KIM, N., CHOI, S., KIM, S. M., LEE, A. C., IM, K., PARK, H. S., KIM, J. A., KIM, K., KIM, I., CHANG, Y. H. & LEE, D. S. 2019. Monozygotic twins with shared de novo GATA2 mutation but dissimilar phenotypes due to differential promoter methylation. *Leuk Lymphoma,* 60**,** 1053-1061.

KIRWAN, M., VULLIAMY, T., MARRONE, A., WALNE, A. J., BESWICK, R., HILLMEN, P., KELLY, R., STEWART, A., BOWEN, D., SCHONLAND, S. O., WHITTLE, A. M., MCVERRY, A., GILLEECE, M. & DOKAL, I. 2009. Defining the pathogenic role of telomerase mutations in myelodysplastic syndrome and acute myeloid leukemia. *Hum Mutat,* 30**,** 1567-73.

KOBAYASHI, S., KOBAYASHI, A., OSAWA, Y., NAGAO, S., TAKANO, K., OKADA, Y., TACHI, N., TERAMOTO, M., KAWAMURA, T., HORIUCHI, T., KATO, S., MAEKAWA, T., YAMAMURA, T., WATANABE, J., HARADA, Y., HARADA, H., SATO, K. & KIMURA, F. 2017. Donor cell leukemia arising from preleukemic clones with a novel germline DDX41 mutation after allogenic hematopoietic stem cell transplantation. *Leukemia,* 31**,** 1020-1022.

KOUZARIDES, T. 2007. SnapShot: Histone-modifying enzymes. *Cell,* 128**,** 802.

KOZYRA, E. J., PASTOR, V., WEHR, C., SAHOO, S. S., VOSS, R., SZVETNIK, E. A., HIRABAYASHI, S., CATALA, A., HASLE, H., HEUVEL-EIBRINK, M. M. V., KALLAY, K., MASETTI, R., MOERLOOSE, B., SCHMUGGE, M., SMITH, O., USSOWICZ, M., STARY, J., MEJSTRIKOVA, E., PASAULIENĖ, R., BAUMANN, I., GÖHRING, G., SCHLEGELBERGER, B., SALZER, U., LÜBBERT, M., TROMPOUKI, E., NIEMEYER, C. M. & WLODARSKI, M. W. 2017. Systematic Assessment of GATA2 Genetic Variation Reveals the Presence of Novel Disease-Causing Synonymous Exonic Mutations. *American Society of Hematology (ASH).* Blood.

LAPPALAINEN, T., MONTGOMERY, S. B., NICA, A. C. & DERMITZAKIS, E. T. 2011. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet,* 89**,** 459-63.

LEFEVRE, S. H., CHAUVEINC, L., STOPPA-LYONNET, D., MICHON, J., LUMBROSO, L., BERTHET, P., FRAPPAZ, D., DUTRILLAUX, B., CHEVILLARD, S. & MALFOY, B. 2002. A T to C mutation in the polypyrimidine tract of the exon 9 splicing site of the RB1 gene responsible for low penetrance hereditary retinoblastoma. *J Med Genet,* 39**,** E21.

LEONE, G., VOSO, M. T., TEOFILI, L. & LUBBERT, M. 2003. Inhibitors of DNA methylation in the treatment of hematological malignancies and MDS. *Clin Immunol,* 109**,** 89-102.

LEWINSOHN, M., BROWN, A. L., WEINEL, L. M., PHUNG, C., RAFIDI, G., LEE, M. K., SCHREIBER, A. W., FENG, J., BABIC, M., CHONG, C. E., LEE, Y., YONG, A., SUTHERS, G. K., POPLAWSKI, N., ALTREE, M., PHILLIPS, K., JAENSCH, L., FINE, M., D'ANDREA, R. J., LEWIS, I. D., MEDEIROS, B. C., POLLYEA, D. A., KING, M. C., WALSH, T., KEEL, S., SHIMAMURA, A., GODLEY, L. A., HAHN, C. N., CHURPEK, J. E. & SCOTT, H. S. 2016. Novel germ line DDX41 mutations define families with a lower age of MDS/AML onset and lymphoid malignancies. *Blood,* 127**,** 1017-23.

LEY, T. J., DING, L., WALTER, M. J., MCLELLAN, M. D., LAMPRECHT, T., LARSON, D. E., KANDOTH, C., PAYTON, J. E., BATY, J., WELCH, J., HARRIS, C. C., LICHTI, C. F., TOWNSEND, R. R., FULTON, R. S., DOOLING, D. J., KOBOLDT, D. C., SCHMIDT, H., ZHANG, Q., OSBORNE, J. R., LIN, L., O'LAUGHLIN, M., MCMICHAEL, J. F., DELEHAUNTY, K. D., MCGRATH, S. D., FULTON, L. A., MAGRINI, V. J., VICKERY, T. L., HUNDAL, J., COOK, L. L., CONYERS, J. J., SWIFT, G. W., REED, J. P., ALLDREDGE, P. A., WYLIE, T., WALKER, J., KALICKI, J., WATSON, M. A., HEATH, S., SHANNON, W. D., VARGHESE, N., NAGARAJAN, R., WESTERVELT, P., TOMASSON, M. H., LINK, D. C., GRAUBERT, T. A., DIPERSIO, J. F., MARDIS, E. R. & WILSON, R. K. 2010. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med,* 363**,** 2424-33.

LEY, T. J., MARDIS, E. R., DING, L., FULTON, B., MCLELLAN, M. D., CHEN, K., DOOLING, D., DUNFORD-SHORE, B. H., MCGRATH, S., HICKENBOTHAM, M., COOK, L., ABBOTT, R., LARSON, D. E., KOBOLDT, D. C., POHL, C., SMITH, S., HAWKINS, A., ABBOTT, S., LOCKE, D., HILLIER, L. W., MINER, T., FULTON, L., MAGRINI, V., WYLIE, T., GLASSCOCK, J., CONYERS, J., SANDER, N., SHI, X., OSBORNE, J. R., MINX, P., GORDON, D., CHINWALLA, A., ZHAO, Y., RIES, R. E., PAYTON, J. E., WESTERVELT, P., TOMASSON, M. H., WATSON, M., BATY, J., IVANOVICH, J., HEATH, S., SHANNON, W. D., NAGARAJAN, R., WALTER, M. J., LINK, D. C., GRAUBERT, T. A., DIPERSIO, J. F. & WILSON, R. K. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature,* 456**,** 66-72.

LIEW, E. & OWEN, C. 2011. Familial myelodysplastic syndromes: a review of the literature. *Haematologica,* 96**,** 1536-42.

LIN, T. C., HOU, H. A., CHOU, W. C., OU, D. L., YU, S. L., TIEN, H. F. & LIN, L. I. 2011. CEBPA methylation as a prognostic biomarker in patients with de novo acute myeloid leukemia. *Leukemia,* 25**,** 32-40.

LIST, A., KURTIN, S., ROE, D. J., BURESH, A., MAHADEVAN, D., FUCHS, D., RIMSZA, L., HEATON, R., KNIGHT, R. & ZELDIS, J. B. 2005. Efficacy of lenalidomide in myelodysplastic syndromes. *N Engl J Med,* 352**,** 549-57.

LISTER, R., PELIZZOLA, M., DOWEN, R. H., HAWKINS, R. D., HON, G., TONTI-FILIPPINI, J., NERY, J. R., LEE, L., YE, Z., NGO, Q. M., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* 462**,** 315-22.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods,* 25**,** 402-8.

LUGER, K., RECHSTEINER, T. J., FLAUS, A. J., WAYE, M. M. & RICHMOND, T. J. 1997. Characterization of nucleosome core particles containing histone proteins made in bacteria. *J Mol Biol,* 272**,** 301-11.

LUNTER, G. & GOODSON, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res,* 21**,** 936-9.

MAKISHIMA, H., YOSHIDA, K., NGUYEN, N., PRZYCHODZEN, B., SANADA, M., OKUNO, Y., NG, K. P., GUDMUNDSSON, K. O., VISHWAKARMA, B. A., JEREZ, A., GOMEZ-SEGUI, I., TAKAHASHI, M., SHIRAISHI, Y., NAGATA, Y., GUINTA, K., MORI, H., SEKERES, M. A., CHIBA, K., TANAKA, H., MURAMATSU, H., SAKAGUCHI, H., PAQUETTE, R. L., MCDEVITT, M. A., KOJIMA, S., SAUNTHARARAJAH, Y., MIYANO, S., SHIH, L. Y., DU, Y., OGAWA, S. & MACIEJEWSKI, J. P. 2013. Somatic SETBP1 mutations in myeloid malignancies. *Nat Genet,* 45**,** 942-6.

MARCUCCI, G., HAFERLACH, T. & DOHNER, H. 2011. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J Clin Oncol,* 29**,** 475-86.

MCNEIL, S. M., NOVELLETTO, A., SRINIDHI, J., BARNES, G., KORNBLUTH, I., ALTHERR, M. R., WASMUTH, J. J., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 1997. Reduced penetrance of the Huntington's disease mutation. *Hum Mol Genet,* 6**,** 775-9.

MCREYNOLDS, L. J., YANG, Y., YUEN WONG, H., TANG, J., ZHANG, Y., MULE, M. P., DAUB, J., PALMER, C., FORURAGHI, L., LIU, Q., ZHU, J., WANG, W., WEST, R. R., YOHE, M. E., HSU, A. P., HICKSTEIN, D. D., TOWNSLEY, D. M., HOLLAND, S. M., CALVO, K. R. & HOURIGAN, C. S. 2019. MDS-associated mutations in germline GATA2 mutated patients with hematologic manifestations. *Leuk Res,* 76**,** 70-75.

MESSEGUER, X., ESCUDERO, R., FARRE, D., NUNEZ, O., MARTINEZ, J. & ALBA, M. M. 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics,* 18**,** 333-4.

METZELER, K. H., HEROLD, T., ROTHENBERG-THURLEY, M., AMLER, S., SAUERLAND, M. C., GORLICH, D., SCHNEIDER, S., KONSTANDIN, N. P., DUFOUR, A., BRAUNDL, K., KSIENZYK, B., ZELLMEIER, E., HARTMANN, L., GREIF, P. A., FIEGL, M., SUBKLEWE, M., BOHLANDER, S. K., KRUG, U., FALDUM, A., BERDEL, W. E., WORMANN, B., BUCHNER, T., HIDDEMANN, W., BRAESS, J., SPIEKERMANN, K. & GROUP, A. S. 2016. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood,* 128**,** 686-98.

MICHAUD, J., WU, F., OSATO, M., COTTLES, G. M., YANAGIDA, M., ASOU, N., SHIGESADA, K., ITO, Y., BENSON, K. F., RASKIND, W. H., ROSSIER, C., ANTONARAKIS, S. E., ISRAELS, S., MCNICOL, A., WEISS, H., HORWITZ, M. & SCOTT, H. S. 2002. In vitro analyses of known and novel RUNX1/AML1 mutations in dominant familial platelet disorder with predisposition to acute myelogenous leukemia: implications for mechanisms of pathogenesis. *Blood,* 99**,** 1364-72.

MICOL, J. B. & ABDEL-WAHAB, O. 2014. Collaborating constitutive and somatic genetic events in myeloid malignancies: ASXL1 mutations in patients with germline GATA2 mutations. *Haematologica,* 99**,** 201-3.

MUKHERJEE, S. 2016. *The Gene: An Intimate History,* New York, Scribner.

NAGATA, Y., NARUMI, S., GUAN, Y., PRZYCHODZEN, B. P., HIRSCH, C. M., MAKISHIMA, H., SHIMA, H., ALY, M., PASTOR, V., KUZMANOVIC, T., RADIVOYEVITCH, T., ADEMA, V., AWADA, H., YOSHIDA, K., LI, S., SOLE, F., HANNA, R., JHA, B. K., LAFRAMBOISE, T., OGAWA, S., SEKERES, M. A., WLODARSKI, M. W., CAMMENGA, J. & MACIEJEWSKI, J. P. 2018. Germline loss-of-function SAMD9 and SAMD9L alterations in adult myelodysplastic syndromes. *Blood,* 132**,** 2309-2313.

NARUMI, S., AMANO, N., ISHII, T., KATSUMATA, N., MUROYA, K., ADACHI, M., TOYOSHIMA, K., TANAKA, Y., FUKUZAWA, R., MIYAKO, K., KINJO, S., OHGA, S., IHARA, K., INOUE, H., KINJO, T., HARA, T., KOHNO, M., YAMADA, S., URANO, H., KITAGAWA, Y., TSUGAWA, K., HIGA, A., MIYAWAKI, M., OKUTANI, T., KIZAKI, Z., HAMADA, H., KIHARA, M., SHIGA, K., YAMAGUCHI, T., KENMOCHI, M., KITAJIMA, H., FUKAMI, M., SHIMIZU, A., KUDOH, J., SHIBATA, S., OKANO, H., MIYAKE, N., MATSUMOTO, N. & HASEGAWA, T. 2016. SAMD9 mutations cause a novel multisystem disorder, MIRAGE syndrome, and are associated with loss of chromosome 7. *Nat Genet,* 48**,** 792-7.

NAZHA, A., AL-ISSA, K., HAMILTON, B. K., RADIVOYEVITCH, T., GERDS, A. T., MUKHERJEE, S., ADEMA, V., ZARZOUR, A., ABUHADRA, N., PATEL, B. J., HIRSCH, C. M., ADVANI, A., PRZYCHODZEN, B., CARRAWAY, H. E., MACIEJEWSKI, J. P. & SEKERES, M. A. 2017. Adding molecular data to prognostic models can improve predictive power in treated patients with myelodysplastic syndromes. *Leukemia,* 31**,** 2848-2850.

NERI, F., INCARNATO, D., KREPELOVA, A., RAPELLI, S., PAGNANI, A., ZECCHINA, R., PARLATO, C. & OLIVIERO, S. 2013. Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol,* 14**,** R91.

NERLOV, C. 2007. The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. *Trends Cell Biol,* 17**,** 318-24.

NICKELS, E. M., SOODALTER, J., CHURPEK, J. E. & GODLEY, L. A. 2013. Recognizing familial myeloid leukemia in adults. *Ther Adv Hematol,* 4**,** 254-69.

NIEMEYER, C. M. & MECUCCI, C. 2017. Practical considerations for diagnosis and management of patients and carriers. *Semin Hematol,* 54**,** 69-74.

NOETZLI, L., LO, R. W., LEE-SHERICK, A. B., CALLAGHAN, M., NORIS, P., SAVOIA, A., RAJPURKAR, M., JONES, K., GOWAN, K., BALDUINI, C., PECCI, A., GNAN, C., DE ROCCO, D., DOUBEK, M., LI, L., LU, L., LEUNG, R., LANDOLT-MARTICORENA, C., HUNGER, S., HELLER, P., GUTIERREZ-HARTMANN, A., XIAYUAN, L., PLUTHERO, F. G., ROWLEY, J. W., WEYRICH, A. S., KAHR, W. H. A., PORTER, C. C. & DI PAOLA, J. 2015. Germline mutations in ETV6 are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. *Nat Genet,* 47**,** 535-538.

NORIS, P., FAVIER, R., ALESSI, M. C., GEDDIS, A. E., KUNISHIMA, S., HELLER, P. G., GIORDANO, P., NIEDERHOFFER, K. Y., BUSSEL, J. B., PODDA, G. M., VIANELLI, N., KERSSEBOOM, R., PECCI, A., GNAN, C., MARCONI, C., AUVRIGNON, A., COHEN, W., YU, J. C., IGUCHI, A., MILLER IMAHIYEROBO, A., BOEHLEN, F., GHALLOUSSI, D., DE ROCCO, D., MAGINI, P., CIVASCHI, E., BIINO, G., SERI, M., SAVOIA, A. & BALDUINI, C. L. 2013. ANKRD26-related thrombocytopenia and myeloid malignancies. *Blood,* 122**,** 1987-9.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

NTZIACHRISTOS, P., ABDEL-WAHAB, O. & AIFANTIS, I. 2016. Emerging concepts of epigenetic dysregulation in hematological malignancies. *Nat Immunol,* 17**,** 1016-24.

OBROCHTA, E. & GODLEY, L. A. 2018. Identifying patients with genetic predisposition to acute myeloid leukemia. *Best Pract Res Clin Haematol,* 31**,** 373-378.

OKANO, M., BELL, D. W., HABER, D. A. & LI, E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell,* 99**,** 247-57.

OMMEM, H. 2016. Monitoring minimal residual disease in acute myeloid leukaemia: a review of the current evolving strategies. *Ther Adv Hematol.,* 1**,** 3–16.

OSSENKOPPELE, G. & SCHUURHUIS, G. J. 2016. MRD in AML: does it already guide therapy decision-making? *Hematology Am Soc Hematol Educ Program,* 2016**,** 356-365.

OSTERGAARD, P., SIMPSON, M. A., CONNELL, F. C., STEWARD, C. G., BRICE, G., WOOLLARD, W. J., DAFOU, D., KILO, T., SMITHSON, S., LUNT, P., MURDAY, V. A., HODGSON, S., KEENAN, R., PILZ, D. T., MARTINEZ-CORRAL, I., MAKINEN, T., MORTIMER, P. S., JEFFERY, S., TREMBATH, R. C. & MANSOUR, S. 2011. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nat Genet,* 43**,** 929-31.

OUYANG, J., GOSWAMI, M., PENG, J., ZUO, Z., DAVER, N., BORTHAKUR, G., TANG, G., MEDEIROS, L. J., JORGENSEN, J. L., RAVANDI, F. & WANG, S. A. 2016. Comparison of Multiparameter Flow Cytometry Immunophenotypic Analysis and Quantitative RT-PCR for the Detection of Minimal Residual Disease of Core Binding Factor Acute Myeloid Leukemia. *Am J Clin Pathol,* 145**,** 769-77.

OWEN, C. J., TOZE, C. L., KOOCHIN, A., FORREST, D. L., SMITH, C. A., STEVENS, J. M., JACKSON, S. C., POON, M. C., SINCLAIR, G. D., LEBER, B., JOHNSON, P. R., MACHETA, A., YIN, J. A., BARNETT, M. J., LISTER, T. A. & FITZGIBBON, J. 2008. Five new pedigrees with inherited RUNX1 mutations causing familial platelet disorder with propensity to myeloid malignancy. *Blood,* 112**,** 4639-45.

PABST, T., EYHOLZER, M., HAEFLIGER, S., SCHARDT, J. & MUELLER, B. U. 2008. Somatic CEBPA mutations are a frequent second event in families with germline CEBPA mutations and familial acute myeloid leukemia. *J Clin Oncol,* 26**,** 5088-93.

PAN, X., MINEGISHI, N., HARIGAE, H., YAMAGIWA, H., MINEGISHI, M., AKINE, Y. & YAMAMOTO, M. 2000. Identification of human GATA-2 gene distal IS exon and its expression in hematopoietic stem cell fractions. *J Biochem,* 127**,** 105-12.

PAPAEMMANUIL, E., GERSTUNG, M., BULLINGER, L., GAIDZIK, V. I., PASCHKA, P., ROBERTS, N. D., POTTER, N. E., HEUSER, M., THOL, F., BOLLI, N., GUNDEM, G., VAN LOO, P., MARTINCORENA, I., GANLY, P., MUDIE, L., MCLAREN, S., O'MEARA, S., RAINE, K., JONES, D. R., TEAGUE, J. W., BUTLER, A. P., GREAVES, M. F., GANSER, A., DÖHNER, K., SCHLENK, R. F., DÖHNER, H. & CAMPBELL, P. J. 2016. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine,* 374**,** 2209-2221.

PASCHKA, P., MARCUCCI, G., RUPPERT, A. S., MROZEK, K., CHEN, H., KITTLES, R. A., VUKOSAVLJEVIC, T., PERROTTI, D., VARDIMAN, J. W., CARROLL, A. J., KOLITZ, J. E., LARSON, R. A., BLOOMFIELD, C. D., CANCER & LEUKEMIA GROUP, B. 2006. Adverse prognostic significance of KIT mutations in adult acute myeloid leukemia with inv(16) and t(8;21): a Cancer and Leukemia Group B Study. *J Clin Oncol,* 24**,** 3904-11.

PASQUET, M., BELLANNE-CHANTELOT, C., TAVITIAN, S., PRADE, N., BEAUPAIN, B., LAROCHELLE, O., PETIT, A., ROHRLICH, P., FERRAND, C., VAN DEN NESTE, E., POIREL, H. A., LAMY, T., OUACHEE-CHARDIN, M., MANSAT-DE MAS, V., CORRE, J., RECHER, C., PLAT, G., BACHELERIE, F., DONADIEU, J. & DELABESSE, E. 2013. High frequency of GATA2 mutations in patients with mild chronic neutropenia evolving to MonoMac syndrome, myelodysplasia, and acute myeloid leukemia. *Blood,* 121**,** 822-9.

PASTOR, V., HIRABAYASHI, S., KAROW, A., WEHRLE, J., KOZYRA, E. J., NIENHOLD, R., RUZAIKE, G., LEBRECHT, D., YOSHIMI, A., NIEWISCH, M., RIPPERGER, T., GOHRING, G., BAUMANN, I., SCHWARZ, S., STRAHM, B., FLOTHO, C., SKODA, R. C., NIEMEYER, C. M. & WLODARSKI, M. W. 2017. Mutational landscape in children with myelodysplastic syndromes is distinct from adults: specific somatic drivers and novel germline variants. *Leukemia,* 31**,** 759-762.

PASTOR, V. B., SAHOO, S. S., BOKLAN, J., SCHWABE, G. C., SARIBEYOGLU, E., STRAHM, B., LEBRECHT, D., VOSS, M., BRYCESON, Y. T., ERLACHER, M., EHNINGER, G., NIEWISCH, M., SCHLEGELBERGER, B., BAUMANN, I., ACHERMANN, J. C., SHIMAMURA, A., HOCHREIN, J., TEDGARD, U., NILSSON, L., HASLE, H., BOERRIES, M., BUSCH, H., NIEMEYER, C. M. & WLODARSKI, M. W. 2018. Constitutional SAMD9L mutations cause familial myelodysplastic syndrome and transient monosomy 7. *Haematologica,* 103**,** 427-437.

PATHAK, A., SEIPEL, K., PEMOV, A., DEWAN, R., BROWN, C., RAVICHANDRAN, S., LUKE, B. T., MALASKY, M., SUMAN, S., YEAGER, M., LABORATORY, N. D. C. G. R., GROUP, N. D. C. S. W., GATTI, R. A., CAPORASO, N. E., MULVIHILL, J. J., GOLDIN, L. R., PABST, T., MCMASTER, M. L. & STEWART, D. R. 2016. Whole exome sequencing reveals a C-terminal germline variant in CEBPA-associated acute myeloid leukemia: 45-year follow up of a large family. *Haematologica,* 101**,** 846-52.

PLON, S. E., ECCLES, D. M., EASTON, D., FOULKES, W. D., GENUARDI, M., GREENBLATT, M. S., HOGERVORST, F. B., HOOGERBRUGGE, N., SPURDLE, A. B., TAVTIGIAN, S. V. & GROUP, I. U. G. V. W. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat,* 29**,** 1282-91.

POLPRASERT, C., SCHULZE, I., SEKERES, M. A., MAKISHIMA, H., PRZYCHODZEN, B., HOSONO, N., SINGH, J., PADGETT, R. A., GU, X., PHILLIPS, J. G., CLEMENTE, M., PARKER, Y., LINDNER, D., DIENES, B., JANKOWSKY, E., SAUNTHARARAJAH, Y., DU, Y., OAKLEY, K., NGUYEN, N., MUKHERJEE, S., PABST, C., GODLEY, L. A., CHURPEK, J. E., POLLYEA, D. A., KRUG, U., BERDEL, W. E., KLEIN, H. U., DUGAS, M., SHIRAISHI, Y., CHIBA, K., TANAKA, H., MIYANO, S., YOSHIDA, K., OGAWA, S., MULLER-TIDOW, C. & MACIEJEWSKI, J. P. 2015. Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell,* 27**,** 658-70.

PREUDHOMME, C., RENNEVILLE, A., BOURDON, V., PHILIPPE, N., ROCHE-LESTIENNE, C., BOISSEL, N., DHEDIN, N., ANDRE, J. M., CORNILLET-LEFEBVRE, P., BARUCHEL, A., MOZZICONACCI, M. J. & SOBOL, H. 2009. High frequency of RUNX1 biallelic alteration in acute myeloid leukemia secondary to familial platelet disorder. *Blood,* 113**,** 5583-7.

RAVANDI, F., WALTER, R. B. & FREEMAN, S. D. 2018. Evaluating measurable residual disease in acute myeloid leukemia. *Blood Adv,* 2**,** 1356-1366.

REINIUS, B. & SANDBERG, R. 2015. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat Rev Genet,* 16**,** 653-64.

RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K., REHM, H. L. & COMMITTEE, A. L. Q. A. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med,* 17**,** 405-24.

RIO-MACHIN, A., CARDOSO, S., ELLISON, A., TAWANA, K., WANG, J., CHELALA, C., PLAGNOL, V., PONTIKOS, N., PAGE, P., REAY, K., WALLIS, Y., RYAN, G., AL SERAIHI, A., WALNE, A., TUMMALA, H., FITZGIBBON, J., DOKAL, I. & VULLIAMY, T. 2018a. Identifying new disease genes in familial myelodysplasia/acute myeloid leukemia *European Hematology Association (EHA).* Hematologica.

RIO-MACHIN, A., TAWANA, K. & FITZGIBBON, J. 2018b. Inherited predisposition to MDS/AML. *HemaSphere - Open access journal of the European Hematology Association,* 2 12-14.

RIPPERGER, T., HOFMANN, W., KOCH, J. C., SHIRNESHAN, K., HAASE, D., WULF, G., ISSING, P. R., KARNEBOGEN, M., SCHMIDT, G., AUBER, B., SCHLEGELBERGER, B., ILLIG, T., ZIRN, B. & STEINEMANN, D. 2018. MDS1 and EVI1 complex locus (MECOM): a novel candidate gene for hereditary hematological malignancies. *Haematologica,* 103**,** e55-e58.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

ROCKOVA, V., ABBAS, S., WOUTERS, B. J., ERPELINCK, C. A., BEVERLOO, H. B., DELWEL, R., VAN PUTTEN, W. L., LOWENBERG, B. & VALK, P. J. 2011. Risk stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers. *Blood,* 118**,** 1069-76.

RODRIGUES, N. P., JANZEN, V., FORKERT, R., DOMBKOWSKI, D. M., BOYD, A. S., ORKIN, S. H., ENVER, T., VYAS, P. & SCADDEN, D. T. 2005. Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood,* 106**,** 477-84.

RODRIGUES, N. P., TIPPING, A. J., WANG, Z. & ENVER, T. 2012. GATA-2 mediated regulation of normal hematopoietic stem/progenitor cell function, myelodysplasia and myeloid leukemia. *Int J Biochem Cell Biol,* 44**,** 457-60.

ROLOFF, G., LAI, C., HOURIGAN, C. & DILLON, L. 2017. Technical Advances in the Measurement of Residual Disease in Acute Myeloid Leukemia. *Journal of Clinical Medicine,* 9**,** 87.

ROMANA, S. P., POIREL, H., LECONIAT, M., FLEXOR, M. A., MAUCHAUFFE, M., JONVEAUX, P., MACINTYRE, E. A., BERGER, R. & BERNARD, O. A. 1995. High frequency of t(12;21) in childhood B-lineage acute lymphoblastic leukemia. *Blood,* 86**,** 4263-9.

ROSE, N. R. & KLOSE, R. J. 2014. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta,* 1839**,** 1362-72.

ROWLEY, J. D. 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature,* 243**,** 290-3.

ROWLEY, J. D. 1980. Chromosome changes in acute leukaemia. *Br J Haematol,* 44**,** 339-46.

SAIDA, S., UMEDA, K., YASUMI, T., MATSUMOTO, A., KATO, I., HIRAMATSU, H., OHARA, O., HEIKE, T. & ADACHI, S. 2016. Successful reduced-intensity stem cell transplantation for GATA2 deficiency before progression of advanced MDS. *Pediatr Transplant,* 20**,** 333-6.

SAKURAI, M., KASAHARA, H., YOSHIDA, K., YOSHIMI, A., KUNIMOTO, H., WATANABE, N., SHIRAISHI, Y., CHIBA, K., TANAKA, H., HARADA, Y., HARADA, H., KAWAKITA, T., KUROKAWA, M., MIYANO, S., TAKAHASHI, S., OGAWA, S., OKAMOTO, S. & NAKAJIMA, H. 2016. Genetic basis of myeloid transformation in familial platelet disorder/acute myeloid leukemia patients with haploinsufficient RUNX1 allele. *Blood Cancer J,* 6**,** e392.

SANDERS, M. A., CHEW, E., FLENSBURG, C., ZEILEMAKER, A., MILLER, S. E., AL HINAI, A. S., BAJEL, A., LUIKEN, B., RIJKEN, M., MCLENNAN, T., HOOGENBOEZEM, R. M., KAVELAARS, F. G., FROHLING, S., BLEWITT, M. E., BINDELS, E. M., ALEXANDER, W. S., LOWENBERG, B., ROBERTS, A. W., VALK, P. J. M. & MAJEWSKI, I. J. 2018. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood,* 132**,** 1526-1534.

SAVAGE, S. A., GIRI, N., BAERLOCHER, G. M., ORR, N., LANSDORP, P. M. & ALTER, B. P. 2008. TINF2, a component of the shelterin telomere protection complex, is mutated in dyskeratosis congenita. *Am J Hum Genet,* 82**,** 501-9.

SCHLENK, R. F., DOHNER, K., KRAUTER, J., FROHLING, S., CORBACIOGLU, A., BULLINGER, L., HABDANK, M., SPATH, D., MORGAN, M., BENNER, A., SCHLEGELBERGER, B., HEIL, G., GANSER, A., DOHNER, H. & GERMAN-AUSTRIAN ACUTE MYELOID LEUKEMIA STUDY, G. 2008. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *N Engl J Med,* 358**,** 1909-18.

SCHNITTGER, S., KERN, W., TSCHULIK, C., WEISS, T., DICKER, F., FALINI, B., HAFERLACH, C. & HAFERLACH, T. 2009. Minimal residual disease levels assessed by NPM1 mutation-specific RQ-PCR provide important prognostic information in AML. *Blood,* 114**,** 2220-31.

SCOTT, H. S., HAHN, C. N., CARMICHAEL, C. L., WILKINS, E. J., CHONG, C. E., BRAUTIGAN, P. J., LI, X. C., STANKOVIC, M., LIN, M., CARMAGNAC, A., BUTCHER, C. M., FRIEND, K. L., EKERT, P. G., KOK, C. H., BROWN, A. L., LEWIS, I. D., TO, L. B., TIMMS, A. E., STOREK, J., MOORE, S., ALTREE, M., ESCHER, R., BARDY, P. G., SUTHERS, G. K., D'ANDREA, R. J. & HORWITZ, M. S. 2010. GATA2 is a New Predisposition Gene for Familial Myelodysplastic Syndrome (MDS) and Acute Myeloid Leukemia (AML). *American Society of Hematology (ASH).* Blood.

SHAIKH, A. & BHARTIYA, D. 2012. Pluripotent Stem Cells in Bone Marrow and Cord Blood. *In:* MOSCHANDREOU, T. E. (ed.) *Blood Cell - An Overview of Studies in Hematology.* IntechOpen.

SHAWKY, R. M. 2014. Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics* 15**,** 103-11.

SHERF, B., NAVARRO, S., HANNAH, R. & WOOD, K. 1996. Dual-Luciferase® reporter assay: An advanced co-reporter technology integrating firefly and Renilla luciferase assays. *Promega Notes,* 57**,** 2-9.

SHIH, A. H., JIANG, Y., MEYDAN, C., SHANK, K., PANDEY, S., BARREYRO, L., ANTONY-DEBRE, I., VIALE, A., SOCCI, N., SUN, Y., ROBERTSON, A., CAVATORE, M., DE STANCHINA, E., HRICIK, T., RAPAPORT, F., WOODS, B., WEI, C., HATLEN, M., BALJEVIC, M., NIMER, S. D., TALLMAN, M., PAIETTA, E., CIMMINO, L., AIFANTIS, I., STEIDL, U., MASON, C., MELNICK, A. & LEVINE, R. L. 2015. Mutational cooperativity linked to combinatorial epigenetic gain of function in acute myeloid leukemia. *Cancer Cell,* 27**,** 502-15.

SMITH, M. L., CAVENAGH, J. D., LISTER, T. A. & FITZGIBBON, J. 2004. Mutation of CEBPA in familial acute myeloid leukemia. *N Engl J Med,* 351**,** 2403-7.

SONG, W. J., SULLIVAN, M. G., LEGARE, R. D., HUTCHINGS, S., TAN, X., KUFRIN, D., RATAJCZAK, J., RESENDE, I. C., HAWORTH, C., HOCK, R., LOH, M., FELIX, C., ROY, D. C., BUSQUE, L., KURNIT, D., WILLMAN, C., GEWIRTZ, A. M., SPECK, N. A., BUSHWELLER, J. H., LI, F. P., GARDINER, K., PONCZ, M., MARIS, J. M. & GILLILAND, D. G. 1999. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat Genet,* 23**,** 166-75.

SPINNER, M. A., SANCHEZ, L. A., HSU, A. P., SHAW, P. A., ZERBE, C. S., CALVO, K. R., ARTHUR, D. C., GU, W., GOULD, C. M., BREWER, C. C., COWEN, E. W., FREEMAN, A. F., OLIVIER, K. N., UZEL, G., ZELAZNY, A. M., DAUB, J. R., SPALDING, C. D., CLAYPOOL, R. J., GIRI, N. K., ALTER, B. P., MACE, E. M., ORANGE, J. S., CUELLAR-RODRIGUEZ, J., HICKSTEIN, D. D. & HOLLAND, S. M. 2014. GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. *Blood,* 123**,** 809-21.

STATHAM, A. L., ROBINSON, M. D., SONG, J. Z., COOLEN, M. W., STIRZAKER, C. & CLARK, S. J. 2012. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res,* 22**,** 1120-7.

STEIN, E. M., DINARDO, C. D., POLLYEA, D. A., FATHI, A. T., ROBOZ, G. J., ALTMAN, J. K., STONE, R. M., DEANGELO, D. J., LEVINE, R. L., FLINN, I. W., KANTARJIAN, H. M., COLLINS, R., PATEL, M. R., FRANKEL, A. E., STEIN, A., SEKERES, M. A., SWORDS, R. T., MEDEIROS, B. C., WILLEKENS, C., VYAS, P., TOSOLINI, A., XU, Q., KNIGHT, R. D., YEN, K. E., AGRESTA, S., DE BOTTON, S. & TALLMAN, M. S. 2017. Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood,* 130**,** 722-731.

STEIN, E. M. & TALLMAN, M. S. 2015. Mixed lineage rearranged leukaemia: pathogenesis and targeting DOT1L. *Curr Opin Hematol,* 22**,** 92-6.

STERN, J. L., THEODORESCU, D., VOGELSTEIN, B., PAPADOPOULOS, N. & CECH, T. R. 2015. Mutation of the TERT promoter, switch to active chromatin, and monoallelic TERT expression in multiple cancers. *Genes Dev,* 29**,** 2219-24.

STONE, R. M., MANDREKAR, S. J., SANFORD, B. L., LAUMANN, K., GEYER, S., BLOOMFIELD, C. D., THIEDE, C., PRIOR, T. W., DOHNER, K., MARCUCCI, G., LO-COCO, F., KLISOVIC, R. B., WEI, A., SIERRA, J., SANZ, M. A., BRANDWEIN, J. M., DE WITTE, T., NIEDERWIESER, D., APPELBAUM, F. R., MEDEIROS, B. C., TALLMAN, M. S., KRAUTER, J., SCHLENK, R. F., GANSER, A., SERVE, H., EHNINGER, G., AMADORI, S., LARSON, R. A. & DOHNER, H. 2017. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N Engl J Med,* 377**,** 454-464.

STRAHL, B. D. & ALLIS, C. D. 2000. The language of covalent histone modifications. *Nature,* 403**,** 41-5.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SWERDLOW, S. H., CAMPO, E., HARRIS, N.L., JAFFE, E.S., PILERI, S.A., STEIN, H., THIELE, J., VARDIMAN, J.W 2008. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues,* Lyon, France, WHO press.

TAWANA, K. 2015. *The Molecular Investigation of Familial Leukaemia.* PhD, Queen Mary University of London.

TAWANA, K., BÖDÖR, C., CAVENAGH, J., JENNER, M. & FITZGIBBON, J. 2013. The Molecular Pathogenesis of Acute Myeloid Leukemia. *CML Leukemia and Lymphoma,* 21**,** 67-77.

TAWANA, K., DRAZER, M. W. & CHURPEK, J. E. 2018. Universal genetic testing for inherited susceptibility in children and adults with myelodysplastic syndrome and acute myeloid leukemia: are we there yet? *Leukemia,* 32**,** 1482-1492.

TAWANA, K. & FITZGIBBON, J. 2016. Inherited DDX41 mutations: 11 genes and counting. *Blood,* 127**,** 960-1.

TAWANA, K., RIO-MACHIN, A., PREUDHOMME, C. & FITZGIBBON, J. 2017a. Familial CEBPA-mutated acute myeloid leukemia. *Semin Hematol,* 54**,** 87-93.

TAWANA, K., WANG, J., KIRALY, P. A., KALLAY, K., BENYO, G., ZOMBORI, M., CSOMOR, J., AL SERAIHI, A., RIO-MACHIN, A., MATOLCSY, A., CHELALA, C., CAVENAGH, J., FITZGIBBON, J. & BODOR, C. 2017b. Recurrent somatic JAK-STAT pathway variants within a RUNX1-mutated pedigree. *Eur J Hum Genet*.

TAWANA, K., WANG, J., RENNEVILLE, A., BODOR, C., HILLS, R., LOVEDAY, C., SAVIC, A., VAN DELFT, F. W., TRELEAVEN, J., GEORGIADES, P., UGLOW, E., ASOU, N., UIKE, N., DEBELJAK, M., JAZBEC, J., ANCLIFF, P., GALE, R., THOMAS, X., MIALOU, V., DOHNER, K., BULLINGER, L., MUELLER, B., PABST, T., STELLJES, M., SCHLEGELBERGER, B., WOZNIAK, E., IQBAL, S., OKOSUN, J., ARAF, S., FRANK, A. K., LAURIDSEN, F. B., PORSE, B., NERLOV, C., OWEN, C., DOKAL, I., GRIBBEN, J., SMITH, M., PREUDHOMME, C., CHELALA, C., CAVENAGH, J. & FITZGIBBON, J. 2015. Disease evolution and outcomes in familial AML with germline CEBPA mutations. *Blood,* 126**,** 1214-23.

TESI, B., DAVIDSSON, J., VOSS, M., RAHIKKALA, E., HOLMES, T. D., CHIANG, S. C. C., KOMULAINEN-EBRAHIM, J., GORCENCO, S., RUNDBERG NILSSON, A., RIPPERGER, T., KOKKONEN, H., BRYDER, D., FIORETOS, T., HENTER, J. I., MOTTONEN, M., NIINIMAKI, R., NILSSON, L., PRONK, C. J., PUSCHMANN, A., QIAN, H., UUSIMAA, J., MOILANEN, J., TEDGARD, U., CAMMENGA, J. & BRYCESON, Y. T. 2017. Gain-of-function SAMD9L mutations cause a syndrome of cytopenia, immunodeficiency, MDS, and neurological symptoms. *Blood,* 129**,** 2266-2279.

TIAN, X., ZHANG, S., LIU, H. M., ZHANG, Y. B., BLAIR, C. A., MERCOLA, D., SASSONE-CORSI, P. & ZI, X. 2013. Histone lysine-specific methyltransferases and demethylases in carcinogenesis: new targets for cancer therapy and prevention. *Curr Cancer Drug Targets,* 13**,** 558-79.

TOBIASSON, M., MCLORNAN, D. P., KARIMI, M., DIMITRIOU, M., JANSSON, M., BEN AZENKOUD, A., JADERSTEN, M., LINDBERG, G., ABDULKADIR, H., KULASEKARARAJ, A., UNGERSTEDT, J., LENNARTSSON, A., EKWALL, K., MUFTI, G. J. & HELLSTROM-LINDBERG, E. 2016. Mutations in histone modulators are associated with prolonged survival during azacitidine therapy. *Oncotarget,* 7**,** 22103-15.

TOWNSLEY, D. M., DUMITRIU, B. & YOUNG, N. S. 2014. Bone marrow failure and the telomeropathies. *Blood,* 124**,** 2775-83.

TSAI, F. Y., KELLER, G., KUO, F. C., WEISS, M., CHEN, J., ROSENBLATT, M., ALT, F. W. & ORKIN, S. H. 1994. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature,* 371**,** 221-6.

TSAI, F. Y. & ORKIN, S. H. 1997. Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood,* 89**,** 3636-43.

TYNER, J. W., TOGNON, C. E., BOTTOMLY, D., WILMOT, B., KURTZ, S. E., SAVAGE, S. L., LONG, N., SCHULTZ, A. R., TRAER, E., ABEL, M., AGARWAL, A., BLUCHER, A., BORATE, U., BRYANT, J., BURKE, R., CARLOS, A., CARPENTER, R., CARROLL, J., CHANG, B. H., COBLENTZ, C., D'ALMEIDA, A., COOK, R., DANILOV, A., DAO, K. T., DEGNIN, M., DEVINE, D., DIBB, J., EDWARDS, D. K. T., EIDE, C. A., ENGLISH, I., GLOVER, J., HENSON, R., HO, H., JEMAL, A., JOHNSON, K., JOHNSON, R., JUNIO, B., KAEMPF, A., LEONARD, J., LIN, C., LIU, S. Q., LO, P., LORIAUX, M. M., LUTY, S., MACEY, T., MACMANIMAN, J., MARTINEZ, J., MORI, M., NELSON, D., NICHOLS, C., PETERS, J., RAMSDILL, J., ROFELTY, A., SCHUFF, R., SEARLES, R., SEGERDELL, E., SMITH, R. L., SPURGEON, S. E., SWEENEY, T., THAPA, A., VISSER, C., WAGNER, J., WATANABE-SMITH, K., WERTH, K., WOLF, J., WHITE, L., YATES, A., ZHANG, H., COGLE, C. R., COLLINS, R. H., CONNOLLY, D. C., DEININGER, M. W., DRUSBOSKY, L., HOURIGAN, C. S., JORDAN, C. T., KROPF, P., LIN, T. L., MARTINEZ, M. E., MEDEIROS, B. C., PALLAPATI, R. R., POLLYEA, D. A., SWORDS, R. T., WATTS, J. M., WEIR, S. J., WIEST, D. L., WINTERS, R. M., MCWEENEY, S. K. & DRUKER, B. J. 2018. Functional genomic landscape of acute myeloid leukaemia. *Nature,* 562**,** 526-531.

VARDIMAN, J. W., THIELE, J., ARBER, D. A., BRUNNING, R. D., BOROWITZ, M. J., PORWIT, A., HARRIS, N. L., LE BEAU, M. M., HELLSTROM-LINDBERG, E., TEFFERI, A. & BLOOMFIELD, C. D. 2009. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood,* 114**,** 937-51.

VICENTE, C., CONCHILLO, A., GARCIA-SANCHEZ, M. A. & ODERO, M. D. 2012. The role of the GATA2 transcription factor in normal and malignant hematopoiesis. *Crit Rev Oncol Hematol,* 82**,** 1-17.

VULLIAMY, T., MARRONE, A., GOLDMAN, F., DEARLOVE, A., BESSLER, M., MASON, P. J. & DOKAL, I. 2001. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature,* 413**,** 432-5.

VULLIAMY, T., MARRONE, A., SZYDLO, R., WALNE, A., MASON, P. J. & DOKAL, I. 2004. Disease anticipation is associated with progressive telomere shortening in families with dyskeratosis congenita due to mutations in TERC. *Nat Genet,* 36**,** 447-9.

VYTOPIL, M., RICCI, E., DELLO RUSSO, A., HANISCH, F., NEUDECKER, S., ZIERZ, S., RICOTTI, R., DEMAY, L., RICHARD, P., WEHNERT, M., BONNE, G., MERLINI, L. & TONIOLO, D. 2002. Frequent low penetrance mutations in the Lamin A/C gene, causing Emery Dreifuss muscular dystrophy. *Neuromuscul Disord,* 12**,** 958-63.

WADDINGTON, C. 1942. The epigenotype. *Endeavour***,** 18-20.

WALNE, A. J., BHAGAT, T., KIRWAN, M., GITIAUX, C., DESGUERRE, I., LEONARD, N., NOGALES, E., VULLIAMY, T. & DOKAL, I. S. 2013a. Mutations in the telomere capping complex in bone marrow failure and related syndromes. *Haematologica,* 98**,** 334-8.

WALNE, A. J., VULLIAMY, T., KIRWAN, M., PLAGNOL, V. & DOKAL, I. 2013b. Constitutional mutations in RTEL1 cause severe dyskeratosis congenita. *Am J Hum Genet,* 92**,** 448-53.

WALNE, A. J., VULLIAMY, T., MARRONE, A., BESWICK, R., KIRWAN, M., MASUNARI, Y., AL-QURASHI, F. H., ALJURF, M. & DOKAL, I. 2007. Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase-associated protein NOP10. *Hum Mol Genet,* 16**,** 1619-29.

WANG, X., MURAMATSU, H., OKUNO, Y., SAKAGUCHI, H., YOSHIDA, K., KAWASHIMA, N., XU, Y., SHIRAISHI, Y., CHIBA, K., TANAKA, H., SAITO, S., NAKAZAWA, Y., MASUNARI, T., HIROSE, T., ELMAHDI, S., NARITA, A., DOISAKI, S., ISMAEL, O., MAKISHIMA, H., HAMA, A., MIYANO, S., TAKAHASHI, Y., OGAWA, S. & KOJIMA, S. 2015. GATA2 and secondary mutations in familial myelodysplastic syndromes and pediatric myeloid malignancies. *Haematologica,* 100**,** e398-401.

WARTIOVAARA-KAUTTO, U., HIRVONEN, E. A. M., PITKANEN, E., HECKMAN, C., SAARELA, J., KETTUNEN, K., PORKKA, K. & KILPIVAARA, O. 2018. Germline alterations in a consecutive series of acute myeloid leukemia. *Leukemia,* 32**,** 2282-2285.

WEBBER, B. R., IACOVINO, M., CHOI, S. H., TOLAR, J., KYBA, M. & BLAZAR, B. R. 2013. DNA methylation of Runx1 regulatory regions correlates with transition from primitive to definitive hematopoietic potential in vitro and in vivo. *Blood,* 122**,** 2978-86.

WEBER, M., HELLMANN, I., STADLER, M. B., RAMOS, L., PAABO, S., REBHAN, M. & SCHUBELER, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet,* 39**,** 457-66.

WEHR, C., GROTIUS, K., CASADEI, S., BLECKMANN, D., BODE, S. F. N., FRYE, B. C., SEIDL, M., GULSUNER, S., KING, M. C., PERCIVAL, M. B., PRITCHARD, C. C., WALSH, T., WU, D., KEEL, S. & SALZER, U. 2018. A novel disease-causing synonymous exonic mutation in GATA2 affecting RNA splicing. *Blood,* 132**,** 1211-1215.

WELCH, J. S., LEY, T. J., LINK, D. C., MILLER, C. A., LARSON, D. E., KOBOLDT, D. C., WARTMAN, L. D., LAMPRECHT, T. L., LIU, F., XIA, J., KANDOTH, C., FULTON, R. S., MCLELLAN, M. D., DOOLING, D. J., WALLIS, J. W., CHEN, K., HARRIS, C. C., SCHMIDT, H. K., KALICKI-VEIZER, J. M., LU, C., ZHANG, Q., LIN, L., O'LAUGHLIN, M. D., MCMICHAEL, J. F., DELEHAUNTY, K. D., FULTON, L. A., MAGRINI, V. J., MCGRATH, S. D., DEMETER, R. T., VICKERY, T. L., HUNDAL, J., COOK, L. L., SWIFT, G. W., REED, J. P., ALLDREDGE, P. A., WYLIE, T. N., WALKER, J. R., WATSON, M. A., HEATH, S. E., SHANNON, W. D., VARGHESE, N., NAGARAJAN, R., PAYTON, J. E., BATY, J. D., KULKARNI, S., KLCO, J. M., TOMASSON, M. H., WESTERVELT, P., WALTER, M. J., GRAUBERT, T. A., DIPERSIO, J. F., DING, L., MARDIS, E. R. & WILSON, R. K. 2012. The origin and evolution of mutations in acute myeloid leukemia. *Cell,* 150**,** 264-78.

WEST, A. H. & CHURPEK, J. E. 2017. Old and new tools in the clinical diagnosis of inherited bone marrow failure syndromes. *Hematology Am Soc Hematol Educ Program,* 2017**,** 79-87.

WEST, A. H., GODLEY, L. A. & CHURPEK, J. E. 2014a. Familial myelodysplastic syndrome/acute leukemia syndromes: a review and utility for translational investigations. *Ann N Y Acad Sci,* 1310**,** 111-8.

WEST, R. R., HSU, A. P., HOLLAND, S. M., CUELLAR-RODRIGUEZ, J. & HICKSTEIN, D. D. 2014b. Acquired ASXL1 mutations are common in patients with inherited GATA2 mutations and correlate with myeloid transformation. *Haematologica,* 99**,** 276-81.

WILLYARD, C. 2017. An epigenetics gold rush: new controls for gene expression. *Nature,* 542**,** 406-408.

WLODARSKI, M. W., COLLIN, M. & HORWITZ, M. S. 2017. GATA2 deficiency and related myeloid neoplasms. *Semin Hematol,* 54**,** 81-86.

WLODARSKI, M. W., HIRABAYASHI, S., PASTOR, V., STARY, J., HASLE, H., MASETTI, R., DWORZAK, M., SCHMUGGE, M., VAN DEN HEUVEL-EIBRINK, M., USSOWICZ, M., DE MOERLOOSE, B., CATALA, A., SMITH, O. P., SEDLACEK, P., LANKESTER, A. C., ZECCA, M., BORDON, V., MATTHES-MARTIN, S., ABRAHAMSSON, J., KUHL, J. S., SYKORA, K. W., ALBERT, M. H., PRZYCHODZIEN, B., MACIEJEWSKI, J. P., SCHWARZ, S., GOHRING, G., SCHLEGELBERGER, B., CSEH, A., NOELLKE, P., YOSHIMI, A., LOCATELLI, F., BAUMANN, I., STRAHM, B., NIEMEYER, C. M. & EWOG, M. D. S. 2016. Prevalence, clinical characteristics, and prognosis of GATA2-related myelodysplastic syndromes in children and adolescents. *Blood,* 127**,** 1387-97; quiz 1518.

WLODARSKI, M. W. & NIEMEYER, C. M. 2017. Introduction: Genetic syndromes predisposing to myeloid neoplasia. *Semin Hematol,* 54**,** 57-59.

WOUTERS, B. J., JORDA, M. A., KEESHAN, K., LOUWERS, I., ERPELINCK-VERSCHUEREN, C. A., TIELEMANS, D., LANGERAK, A. W., HE, Y., YASHIRO-OHTANI, Y., ZHANG, P., HETHERINGTON, C. J., VERHAAK, R. G., VALK, P. J., LOWENBERG, B., TENEN, D. G., PEAR, W. S. & DELWEL, R. 2007. Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood,* 110**,** 3706-14.

XIAO, H., SHI, J., LUO, Y., TAN, Y., HE, J., XIE, W., ZHANG, L., WANG, Y., LIU, L., WU, K., YU, X., CAI, Z., LIN, M., YE, X. & HUANG, H. 2011. First report of multiple CEBPA mutations contributing to donor origin of leukemia relapse after allogeneic hematopoietic stem cell transplantation. *Blood,* 117**,** 5257-60.

YAMAGUCHI, H., CALADO, R. T., LY, H., KAJIGAYA, S., BAERLOCHER, G. M., CHANOCK, S. J., LANSDORP, P. M. & YOUNG, N. S. 2005. Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. *N Engl J Med,* 352**,** 1413-24.

YATES, J. W., WALLACE, H. J., JR., ELLISON, R. R. & HOLLAND, J. F. 1973. Cytosine arabinoside (NSC-63878) and daunorubicin (NSC-83142) therapy in acute nonlymphocytic leukemia. *Cancer Chemother Rep,* 57**,** 485-8.

YATES, L. R. & CAMPBELL, P. J. 2012. Evolution of the cancer genome. *Nat Rev Genet,* 13**,** 795-806.

YOSHIMI, A., TOYA, T., KAWAZU, M., UENO, T., TSUKAMOTO, A., IIZUKA, H., NAKAGAWA, M., NANNYA, Y., ARAI, S., HARADA, H., USUKI, K., HAYASHI, Y., ITO, E., KIRITO, K., NAKAJIMA, H., ICHIKAWA, M., MANO, H. & KUROKAWA, M. 2014. Recurrent CDC25C mutations drive malignant transformation in FPD/AML. *Nat Commun,* 5**,** 4770.

ZHANG, M. Y., CHURPEK, J. E., KEEL, S. B., WALSH, T., LEE, M. K., LOEB, K. R., GULSUNER, S., PRITCHARD, C. C., SANCHEZ-BONILLA, M., DELROW, J. J., BASOM, R. S., FOROUHAR, M., GYURKOCZA, B., SCHWARTZ, B. S., NEISTADT, B., MARQUEZ, R., MARIANI, C. J., COATS, S. A., HOFMANN, I., LINDSLEY, R. C., WILLIAMS, D. A., ABKOWITZ, J. L., HORWITZ, M. S., KING, M. C., GODLEY, L. A. & SHIMAMURA, A. 2015. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nat Genet,* 47**,** 180-5.

ZHANG, X., LV, D., ZHANG, Y., LIU, Q. & LI, Z. 2016. Clonal evolution of acute myeloid leukemia highlighted by latest genome sequencing studies. *Oncotarget,* 7**,** 58586-58594.

ZHANG, Y. & REINBERG, D. 2001. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev,* 15**,** 2343-60.

Appendix 1. Publication in *Leukemia*

Leukemia
https://doi.org/10.1038/s41375-018-0134-9

BRIEF COMMUNICATION

Acute myeloid leukemia

GATA2 monoallelic expression underlies reduced penetrance in inherited *GATA2*-mutated MDS/AML

Ahad F. Al Seraihi[1] · Ana Rio-Machin[1] · Kiran Tawana[1] · Csaba Bödör[2] · Jun Wang[3] · Ai Nagano[3] · James A. Heward[1] · Sameena Iqbal[1] · Steven Best[4] · Nicholas Lea[4] · Donal McLornan[5] · Emilia J. Kozyra[6,7] · Marcin W. Wlodarski[7] · Charlotte M. Niemeyer[7] · Hamish Scott[8] · Chris Hahn[8] · Alicia Ellison[9] · Hemanth Tummala[9] · Shirleny Romualdo Cardoso[9] · Tom Vulliamy[9] · Inderjeet Dokal[9] · Tom Butler[10] · Matthew Smith[10] · Jamie Cavenagh[10] · Jude Fitzgibbon[1]

Received: 21 December 2017 / Accepted: 4 April 2018

While the majority of myelodysplasia and acute myeloid leukemia (MDS/AML) cases are sporadic, rare familial predisposition syndromes have been delineated and now represent a separate disease entity in the revised World Health Organization (WHO) classification of myeloid neoplasms [1]. Germline mutations in ~14 disease genes have been uncovered thus far, with *GATA2* representing one of the key transcriptional regulators commonly mutated in inherited MDS/AML [2]. Increasing evidence suggests that aberrations in *GATA2* impair its transcription and promoter activation, leading to a loss-of-function, supporting a mechanism of *GATA2* haploinsufficiency [3–5]. Reduced penetrance, the observation that family members carry an identical germline mutation yet display variable clinical manifestations, is common and poses a clinical challenge in the diagnosis and management of familial leukemia's, particularly when identifying "silent" mutation carriers for genetic screening and exclusion as potential stem cell transplant donors [6, 7]. Indeed, we have noted that reduced penetrance is a feature among certain *GATA2*-mutated MDS/AML families [8], especially those harboring missense germline mutations such as c.1061C>T (p.Thr354Met) (Table S1) although the precise molecular explanation of such occurrence has not been investigated.

Analysis of five MDS/AML families harboring p.Thr354Met *GATA2* mutations displayed significant intra- and interfamilial variations in disease latency, phenotype, and penetrance (Figure S1). These observations suggest that individuals require additional co-operating events for the development of overt malignancy within the context of a shared germline mutation. To investigate this hypothesis further, we examined an extensive five-generation pedigree [9] (Fig. 1a) where two first-degree cousins (IV.1 and IV.6) developed high-risk MDS/AML with monosomy 7, while a third cousin (IV.10) presented with recurrent minor infections and significant monocytopenia [0.1×10^9/L] and

These authors contributed equally: Ahad F. Al Seraihi, Ana Rio-Machin.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41375-018-0134-9) contains supplementary material, which is available to authorized users.

✉ Ahad F. Al Seraihi
a.f.h.alseraihi@qmul.ac.uk

✉ Jude Fitzgibbon
j.fitzgibbon@qmul.ac.uk

1 Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, London, UK

2 MTA-SE Lendulet Molecular Oncohematology Research Group, 1st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary

3 Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, London, UK

4 Laboratory for Molecular Haemato-Oncology, King's College Hospital NHS Foundation Trust, London, UK

5 Department of Haematological Medicine, King's College Hospital, London, UK

6 University of Freiburg, Faculty of Biology, Freiburg, Germany

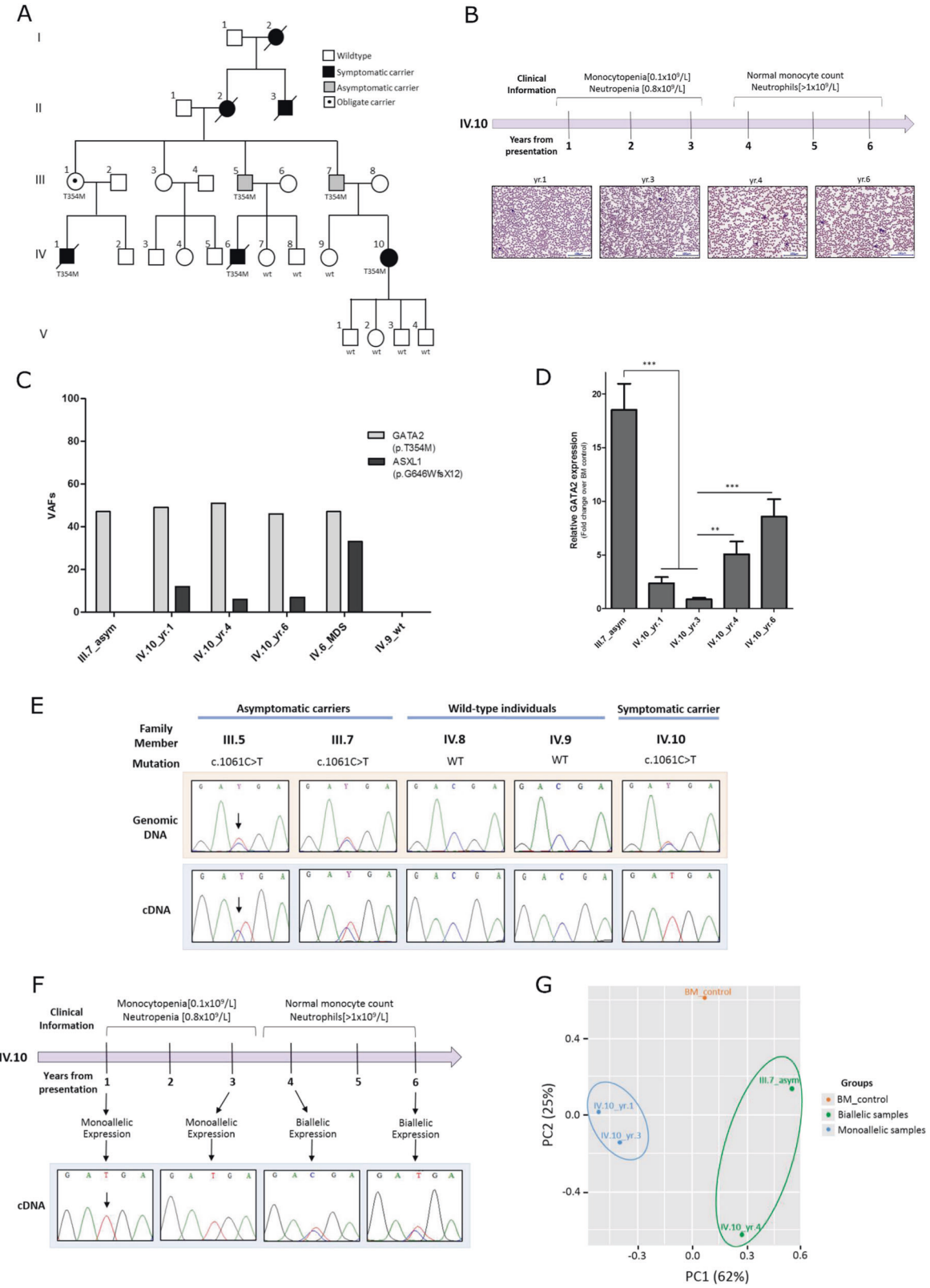
7 Pediatric Hematology and Oncology, University Children's Hospital Freiburg, Freiburg, Germany

8 Centre for Cancer Biology, SA Pathology, University of South Australia, Adelaide, SA, Australia

9 Centre for Genomics and Child Health, Blizard Institute, Queen Mary University of London, London, UK

10 Department of Haemato-Oncology, St. Bartholomew's Hospital, Barts Health NHS Trust, London, UK

Published online: 19 April 2018



neutropenia [0.8×10^9/L] in year (yr.) 1–3 which subsequently stabilized (monocyte count, neutrophils [$>1 \times 10^9$/L]) 3 years after presentation (Fig. 1b). This contrasted with the parental generation (III.1, III.5, and III.7) where mutation carriers remain symptom-free with no evidence of hematopoietic abnormality over 60 years of age.

◀ **Fig. 1** Investigating the molecular mechanisms underlying the reduced penetrance of germline p.Thr354Met mutations observed in a *GATA2*-mutated MDS/AML family. **a** Genogram of the *GATA2*-mutated pedigree. Squares denote males and circles denote females. This five-generation MDS/AML family presented to Barts Health hospital in London with identical germline *GATA2* mutations (p.Thr354Met; c.1061C>T) and variable clinical manifestations. Two first-degree cousins (IV.1 and IV.6) presented at 23 and 18 years of age, respectively, with high-grade MDS transforming to AML and monosomy 7. Both cousins died post allogeneic hematopoietic stem cell transplant (HSCT) due to transplant-related complications (IV.1 from graft vs. host disease (GvHD) and IV.6 from relapsed MDS/AML). Ten years later, their first cousin (IV.10) developed symptoms at 31 years, including recurrent minor infections and significant leukopenia (monocytopenia [0.1×10^9/L] and neutropenia [0.8×10^9/L]) with mild macrocytosis and normal hemoglobin and platelet counts. She remains under close surveillance where her blood counts are routinely monitored. All four of her children have inherited her WT *GATA2* allele. Similarly, members (IV.7, IV.8, and IV.9) were screened for the mutation and all have a WT *GATA2* configuration. The paternal grandmother (II.2) of IV.10 as well as her paternal great-uncle (II.3) and great-grandmother (I.2) all were reported to have died of AML (ages of disease onset were 53, 24, and 53-years old, respectively). Not only did *GATA2* mutations correlate with early age of disease onset in the fourth generation (IV.1/23 yr., IV.6/18 yr., and IV.10/31 yr.), but the parental third-generation carriers (III.1, III.5, and III.7) remain hematologically normal and symptom-free into their mid–late 60s. No material was available from other family members. **b** A clinical timeline of IV.10 showing the change in clinical parameters over the course of disease presentation. Photographs of peripheral blood smears from IV.10 (yr. 1, 3, 4, and 6) stained with May-Grünwald Giemsa staining. Magnification: ×20. **c** Secondary *ASXL1* mutations: variant allele frequencies of *GATA2* germline mutation and *ASXL1* acquired mutation. Samples from three individuals were sequenced: one asymptomatic parent (III.7), one deceased MDS/AML cousin (IV.6), and across three time-points (yr. 1, 4, and 6) from the symptomatic patient (IV.10) reflecting disease evolution. **d** *GATA2* global expression measured by qRT-PCR of bone marrow samples and normalized to healthy bone marrow control: downregulation in IV.10_yr.1 compared with III.7 and downregulation in IV.10_yr.1–3 *GATA2* expression compared with IV.10_yr.4–6. The average of five independent experiments is shown. Statistical significance was determined at $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$ using a *t*-test with Bonferroni correction. Error bars represent standard error of the mean (SEM). **e** *GATA2* monoallelic expression of the mutant allele in symptomatic (IV.10) vs. asymptomatic carriers (III.5 and III.7), as measured by cDNA sequencing of bone marrow samples. **f** Correlation of monoallelic *GATA2* expression with disease symptoms across the time-points studied in IV.10 with reactivation of the WT allele **"C"** expression noted 3 years after presentation, concurrent with improvements in hematological parameters. **g** RNA-seq analysis: principal component analysis (PCA) plot showing a good separation between *GATA2* biallelic (green) and monoallelic (blue) groups based on all transcriptomes

We therefore started with targeted deep sequencing of 33 genes frequently mutated in MDS/AML to define the landscape of secondary genetic mutations across mutation carriers. Notably, while no acquired mutations were detected in asymptomatic family members, all affected cousins analyzed shared an identical somatic *ASXL1* mutation (p.Gly646TrpfsTer12) (Fig. 1c). The variant allele frequency (VAF), however, was lower (12%) in IV.10 and remained stable (range 12–6%) over a 6-year monitoring period. While the co-occurrence of *ASXL1* and *GATA2* mutations has been proposed as one mechanism for driving the onset and severity of disease symptoms [9–11], the low VAF of *ASXL1* mutation and stable improvement in hematopoiesis at IV.10 later follow-up suggested that a combination of *GATA2–ASXL1* mutation alone is insufficient to promote clonal expansion and leukemic transformation, as this secondary somatic hit may not represent disease progression or identify when treatment is indicated. Intriguingly, apart from the *ASXL1* mutation, no other acquired mutations were detected in the 33-myeloid genes assessed in the affected individuals. Moreover, on the basis of our observations and in agreement with previous studies [12, 13], it seems that monosomy 7 in IV.1 and IV.6 is acquired following acquisition of *ASXL1* mutations, hence contributing to the malignancy but not initiating symptoms.

We next considered whether disease symptoms are modulated by endogenous levels of *GATA2*. Quantitative real-time PCR (qRT-PCR) of bone marrow material demonstrated total *GATA2* expression to be significantly lower in the symptomatic (IV.10-yr.1) compared with an asymptomatic carrier (III.7) (Fig. 1d). Significantly, Sanger sequencing of the cDNA template revealed striking allele-specific expression (ASE), favoring the mutant (T) allele with the absence of the wild-type (WT) (C) allele expression in the symptomatic patient (IV.10), contrasting with biallelic expression in asymptomatic members (III.5 and III.7) (Fig. 1e). This observation was validated by cDNA cloning of III.7 and IV.10 bone marrow samples and subsequent Sanger sequencing of individual clones (Figure S2). As this suggested that an allelic imbalance in WT:mutant *GATA2* expression ratio may account for the variable disease penetrance in this pedigree, we assessed *GATA2* expression in IV.10 over a 6-year disease period at four time-points (yr. 1, 3, 4, and 6), demonstrating increased *GATA2* expression at later time-points (yr. 4 and 6) (Fig. 1d) coinciding with reactivation of the WT (C) allele expression (Fig. 1f) and an improvement in hematological parameters, in the absence of any clinical intervention (Fig. 1b).

To test whether monoallelic *GATA2* expression has an impact on the transcriptome driving the onset of disease symptoms, we performed RNA-seq with a view of examining downstream biological features distinctive of *GATA2* monoallelic (IV.10-yr.1 and 3) vs. biallelic (IV.10-yr.4 and III.7) groups. Unsupervised analysis revealed a clear separation between *GATA2* monoallelic and biallelic samples (Fig. 1g, S3 and Table S2). It was noteworthy that certain canonical pathways and gene sets related to tumorigenesis (e.g., DNA replication and cell cycle) were enriched in *GATA2* monoallelic vs. biallelic groups (Figure S4), potentially reflecting the clinical and phenotypic switch

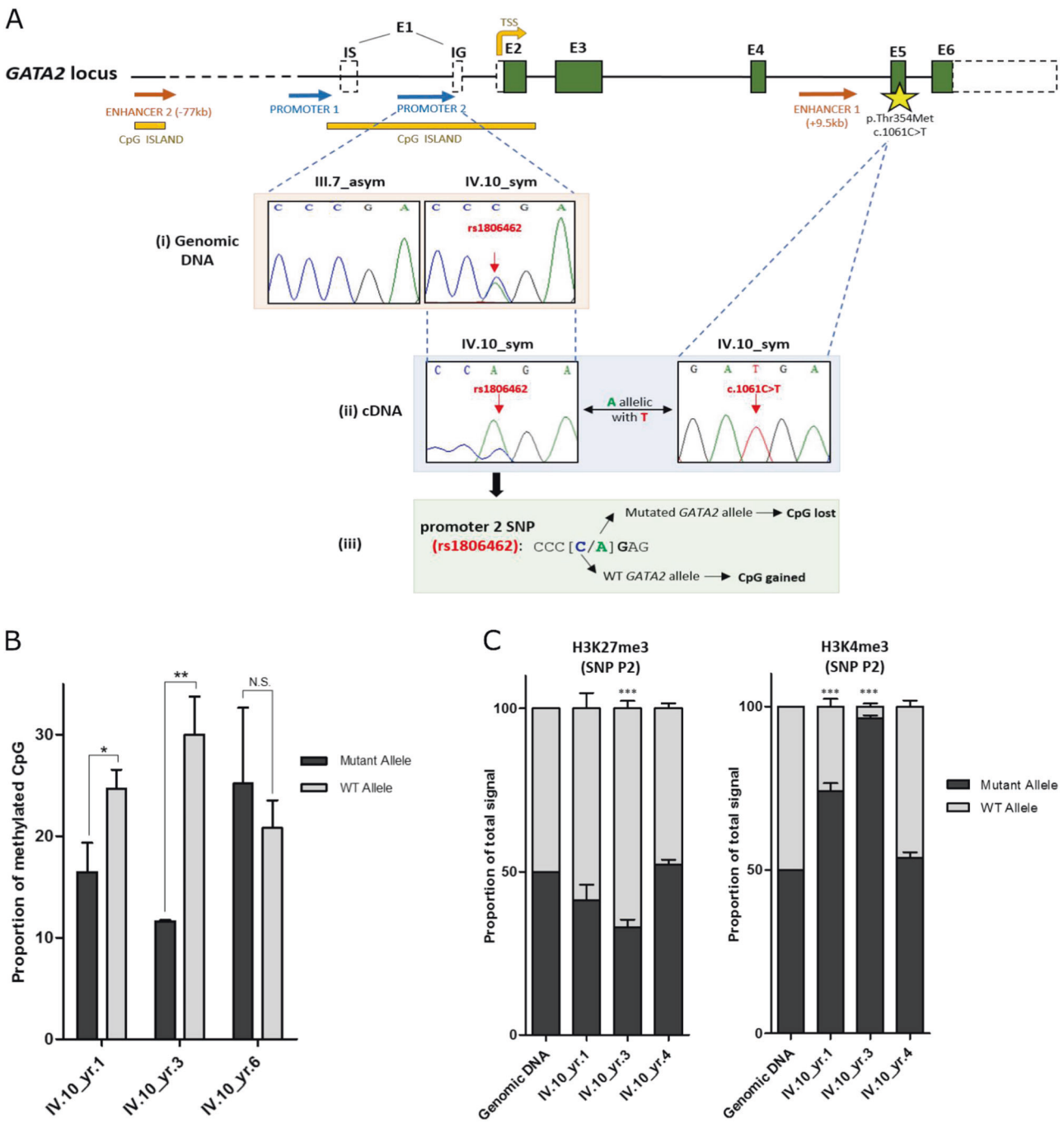


Fig. 2 Elucidating the molecular mechanisms driving allele-specific changes in *GATA2* expression. **a(i)** A noncoding SNP (rs1806462 [C/A]) located within the second *GATA2* promoter region overlapping a CpG island was detected in the symptomatic (IV.10) but not in asymptomatic members (III.7). **a(ii)** Given the location of promoter 2 SNP within the 5'UTR, a haplotype between the SNP allele "A" and the germline mutant allele "T" was established, providing a means of distinguishing between mutant and WT alleles in subsequent experiments. **a(iii)** This promoter SNP [C/A] removes a CpG methylation site in the mutant allele "A" and generates a CpG methylation site in the WT allele "C". **b** The proportion of methylated CpGs between mutant and WT alleles across the three time-points of IV.10. WT allele is significantly more methylated than the mutant allele in monoallelic samples (yr. 1 and yr. 3), whereas no significant allele-specific differences in methylation were observed in a biallelic-expressing sample (yr. 6). The average of three independent experiments is shown. **c** Quantification of mutant and WT allele ChIP sequence peak heights across the time-points of IV.10 based on Sanger sequencing. H3K4me3 activation mark favoring the mutant allele was enriched in monoallelic samples (yr. 1 and yr. 3) compared with the biallelic sample (yr. 4). The average of three independent experiments is shown. Statistical significance was determined at $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$ using a *t*-test with Bonferroni correction. NS corresponds to nonsignificant comparisons. Error bars represent SEM

between these two groups. We also noted a significant overexpression of genes with *GATA2* cofactor *PU.1* motifs in their regulatory regions (p value NES = 2.06) in *GATA2* biallelic vs. monoallelic samples, in support of a recent finding [14] that p.Thr354Met mutants bind and interact with PU.1 more tightly than WT, thus leading to sequestration of PU.1 from its normal cellular functions. Consequently, the transcriptional activation triggered by PU.1 will be diminished in our *GATA2* monoallelic samples.

The differences observed in these gene-expression profiles prompted us to explore the molecular mechanisms underlying monoallelic *GATA2* expression. We hypothesized that these allele-specific changes in *GATA2* expression are driven by transient epigenetic mechanisms that include changes in DNA methylation and chromatin mark deposition. A CpG single-nucleotide polymorphism (CpG-SNP) (rs1806462) [C/A] located within the promoter and 5′UTR of *GATA2* overlapping a CpG island offered a marker to distinguish between mutant and WT alleles where this SNP creates/abolishes a CpG dinucleotide within the *GATA2* promoter region (Fig. 2a). More specifically, cDNA sequencing of 5′UTR allowed us to define haplotypes, where the promoter SNP allele (A) resides on the germline mutant *GATA2* allele (T) (Fig. 2a(ii)). Apart from IV.10, no other family members and only 2/12 individuals from pedigrees presented in Figure S1 were heterozygous for this SNP (one of whom is an asymptomatic carrier). Therefore, we do not infer that this haplotype would contribute to the progression of symptoms. Instead, we used this SNP to determine whether allele-specific differences in DNA methylation could explain the silencing of WT *GATA2* allele expression observed in earlier time-points of IV.10. As illustrated in Fig. 2b and S5, bisulfite sequencing of a 200-bp region encompassing rs1806462 demonstrated a significant increase in promoter methylation in the WT allele of IV.10 in yr. 1 and yr. 3 following diagnosis, in contrast with the absence of allele-specific differences in methylation at a later time-point.

We next sought to establish whether these allele-specific changes in *GATA2* methylation and expression are accompanied by changes in chromatin structure at the promoter. H3K4me3 and H3K27me3 define poised or closed chromatin, respectively, rendering them more or less accessible for transcription factors, thereby regulating gene expression [15]. The deposition of these bivalent marks was assessed in IV.10 by allele-specific chromatin immunoprecipitation (ChIP) followed by Sanger sequencing within *GATA2* promoter region encompassing the SNP rs1806462 [C/A]. While there were no apparent allele-specific differences in H3K27me3 deposition across the different time-points of IV.10, an enrichment in the deposition of H3K4me3 on the promoter of the mutant allele (A) relative to the WT allele (C) was noted in IV.10 monoallelic samples (yr. 1 and 3) (Fig. 2c, S6 and S7). In contrast, and consistent with the pattern observed with DNA methylation, there was no demonstrable difference in H3K4me3 deposition in the IV.10 biallelic sample (yr. 4), coinciding with reactivation of the WT allele expression and an overall improvement in clinical parameters. We believe that these observations are in keeping with the notion that H3K4me3 occupancy inhibits de novo DNA methylation [16] which was borne out by subsequent bisulfite sequencing of H3K4me3-enriched DNA from our ChIP experiments, demonstrating that DNA methylation and H3K4me3 deposition are mutually exclusive in our IV.10 samples (Figure S8).

Collectively, our findings provide a step forward in understanding the molecular mechanisms underlying reduced penetrance in *GATA2*-mutated MDS/AML pedigrees, which may be governed by the acquisition of additional co-operating mutations (e.g., *ASXL1*) combined with dynamic epigenetic reprogramming and subsequent allele-specific expression of *GATA2* mutant allele, adding another level of complexity to the (epi)genetic basis of familial MDS/AML.

Acknowledgements We are indebted to the family investigated in this study whose members have kindly donated samples for research. We also thank all the clinicians who have looked after this family over the years. This study was supported by the Saudi Arabian Ministry of Higher Education through a doctoral scholarship awarded to A.F.A.S. and a Bloodwise Programme grant (14032) awarded to J.F., T.V., and I.D.

Author contributions J.F., A.F.A.S., and A.R.-M. designed the study; A.F.A.S. and A.R.-M. performed the experiments; A.F.A.S., A.R.-M., K.T., and J.F. analyzed the data and wrote the manuscript; K.T., H.S., C.H., T.V., I.D., M.S., and J.C. collated familial clinical information; S.I. provided patient material from tissue bank; S.B., N.L., and D.M. performed targeted deep sequencing; J.W. and A.N. carried out RNA-seq analysis; J.A.H. provided technical ChIP expertise; E.J.K., M.W. W., and C.M.N. provided familial samples; T.B. provided patient blood films; and C.B., A.E., S.R.C., H.T., T.V., and I.D. assisted with data analysis and contributed to the study with fruitful discussions. All authors read, reviewed, and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127:2391–405.
2. Nickels EM, Soodalter J, Churpek JE, Godley LA. Recognizing familial myeloid leukemia in adults. Ther Adv Hematol. 2013;4:254–69.
3. Celton M, Forest A, Gosse G, Lemieux S, Hebert J, Sauvageau G, et al. Epigenetic regulation of GATA2 and its impact on normal karyotype acute myeloid leukemia. Leukemia. 2014;28:1617–26.
4. Cortes-Lavaud X, Landecho MF, Maicas M, Urquiza L, Merino J, Moreno-Miralles I, et al. GATA2 germline mutations impair GATA2 transcription, causing haploinsufficiency: functional analysis of the p.Arg396Gln mutation. J Immunol. 2015;194:2190–8.
5. Hsu AP, Johnson KD, Falcone EL, Sanalkumar R, Sanchez L, Hickstein DD, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. Blood. 2013;121:3830–7. S3831-3837
6. University of Chicago Hematopoietic Malignancies Cancer Risk T. How I diagnose and manage individuals at risk for inherited myeloid malignancies. Blood. 2016;128:1800–13.
7. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132:1077–130.
8. Hahn CN, Chong CE, Carmichael CL, Wilkins EJ, Brautigan PJ, Li XC, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. Nat Genet. 2011;43:1012–7.
9. Bodor C, Renneville A, Smith M, Charazac A, Iqbal S, Etancelin P, et al. Germ-line GATA2 p.THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival. Haematologica. 2012;97:890–4.
10. Boultwood J, Perry J, Pellagatti A, Fernandez-Mercado M, Fernandez-Santamaria C, Calasanz MJ, et al. Frequent mutation of the polycomb-associated gene ASXL1 in the myelodysplastic syndromes and in acute myeloid leukemia. Leukemia. 2010;24:1062–5.
11. West RR, Hsu AP, Holland SM, Cuellar-Rodriguez J, Hickstein DD. Acquired ASXL1 mutations are common in patients with inherited GATA2 mutations and correlate with myeloid transformation. Haematologica. 2014;99:276–81.
12. Wang X, Muramatsu H, Okuno Y, Sakaguchi H, Yoshida K, Kawashima N, et al. GATA2 and secondary mutations in familial myelodysplastic syndromes and pediatric myeloid malignancies. Haematologica. 2015;100:e398–401.
13. Pastor V, Hirabayashi S, Karow A, Wehrle J, Kozyra EJ, Nienhold R, et al. Mutational landscape in children with myelodysplastic syndromes is distinct from adults: specific somatic drivers and novel germline variants. Leukemia. 2017;31:759–62.
14. Chong CE, Venugopal P, Stokes PH, Lee YK, Brautigan PJ, Yeung DTO, et al. Differential effects on gene transcription and hematopoietic differentiation correlate with GATA2 mutant disease phenotypes. Leukemia. 2017;32:194–202.
15. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012;150:12–27.
16. Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. Biochim Biophys Acta. 2014;1839:1362–72.

Appendix 2. Table 1

Targeted deep sequencing results – based on the 33-myeloid gene panel.

Sample	Gene	Variant	Chr.	Coordinate	Amino Acid Change	Transcript Change	VAF%	Variant Classification
IV.10_yr.1	*GATA2*	G>G/A	3	128200744	p.Thr354Met	c.1061C>T	49	Pathogenic
	TET2	A>A/G	4	106196951	p.Ile1762Val	c.5284A>G	51	High freq SNP
	TP53	G>C/C	17	7579472	p.Pro72Arg	c.215C>G	99	High freq SNP
	ASXL1	A>A/AG	20	31022441	p.Gly646TrpfsTer12	c.1926_1927insG	12	Presumed Pathogenic
	ASXL1	T>C/C	20	31022959	p.Leu815Pro	c.2444T>C	100	High freq SNP
IV.10_yr.4	*GATA2*	G>G/A	3	128200744	p.Thr354Met	c.1061C>T	51	Pathogenic
	TET2	A>A/G	4	106196951	p.Ile1762Val	c.5284A>G	52	High freq SNP
	TP53	G>C/C	17	7579472	p.Pro72Arg	c.215C>G	100	High freq SNP
	ASXL1	A>A/AG	20	31022441	p.Gly646TrpfsTer12	c.1926_1927insG	6	Presumed Pathogenic
	ASXL1	T>C/C	20	31022959	p.Leu815Pro	c.2444T>C	100	High freq SNP
IV.10_yr.6	*GATA2*	G>G/A	3	128200744	p.Thr354Met	c.1061C>T	46	Pathogenic
	TET2	A>A/G	4	106196951	p.Ile1762Val	c.5284A>G	51	High freq SNP
	TP53	G>C/C	17	7579472	p.Pro72Arg	c.215C>G	100	High freq SNP
	ASXL1	A>A/AG	20	31022441	p.Gly646TrpfsTer12	c.1926_1927insG	7	Presumed Pathogenic
	ASXL1	T>C/C	20	31022959	p.Leu815Pro	c.2444T>C	100	High freq SNP
III.7_Asym	*GATA2*	G>G/A	3	128200744	p.Thr354Met	c.1061C>T	47	Pathogenic
	TET2	C>C/T	4	106156187	p.Pro363Leu	c.1088C>T	50	Unknown Significance
	TET2	T>T/G	4	106196829	p.Leu1721Trp	c.5162T>G	47	Unknown Significance
	TP53	G>C/C	17	7579472	p.Pro72Arg	c.215C>G	99	High freq SNP
	ASXL1	T>C/C	20	31022959	p.Leu815Pro	c.2444T>C	100	High freq SNP

IV.6_MDS/AML	*GATA2*	G>G/A	3	128200744	p.Thr354Met	c.1061C>T	47	Pathogenic
	TET2	A>G/G	4	106196951	p.Ile1762Val	c.5284A>G	100	High freq SNP
	EZH2	C>C/G	7	148525904	p.Asp185His	c.553G>C	15	High freq SNP
	TP53	G>G/C	17	7579472	p.Pro72Arg	c.215C>G	49	High freq SNP
	ASXL1	A>A/AG	20	31022441	p.Gly646TrpfsTer12	c.1926_1927insG	33	Presumed Pathogenic
	ASXL1	T>C/C	20	31022959	p.Leu815Pro	c.2444T>C	100	High freq SNP

Appendix 3. Table 2

Select differentially expressed (DE) genes between *GATA2* biallelic and monoallelic group

(Top 30 upregulated and top 30 downregulated genes)

Table S2. List of 2,432 differentially expressed (DE) genes between *GATA2* biallelic (green) and monoallelic (blue) groups

								GATA2 Monoallelic group		*GATA2* Biallelic group		
ID	Name	Description	log2FC (Bi_ vs Mono_allelic)	logCPM	LR	PValue	FDR	IV.10_yr.1	IV.10_yr.3	IV.10_yr.4	III.7_asym	BM_control
ENSG00000112077	RHAG	Rh-associated glycoprotein [Source:HGNC Symbol;Acc:10006]	-7.065450382	5.860144216	250.1161537	2.45E-56	4.30E-52	6.93698598	7.148898694	-0.82139248	-1.572251189	4.628306871
ENSG00000107562	CXCL12	chemokine (C-X-C motif) ligand 12 [Source:HGNC Symbol;Acc:10672]	-8.330878237	6.904575843	221.365128	4.56E-50	4.00E-46	8.463534146	7.896112965	-2.073364206	-0.956081001	2.751827722
ENSG00000169877	AHSP	alpha hemoglobin stabilizing protein [Source:HGNC Symbol;Acc:18075]	-7.680771874	6.64991784	218.0163748	2.45E-49	1.43E-45	7.868211149	7.83620084	0.505333341	-1.854339469	5.243726669
ENSG00000163554	SPTA1	spectrin, alpha, erythrocytic 1 (elliptocytosis 2) [Source:HGNC Symbol;Acc:11272]	-6.718271315	6.277401811	211.6474111	6.00E-48	2.64E-44	6.962770807	7.687236572	-0.589913612	0.154058352	5.777640584
ENSG00000223609	HBD	hemoglobin, delta [Source:HGNC Symbol;Acc:4829]	-6.948481557	7.476636825	209.3272706	1.93E-47	6.76E-44	8.65012583	8.713064225	1.855590689	0.984547783	5.982087099
ENSG00000196415	PRTN3	proteinase 3 [Source:HGNC Symbol;Acc:9495]	-6.967317693	6.019308576	198.8586812	3.71E-45	1.08E-41	6.122897312	6.891813193	-2.073364206	-1.133744865	7.075716419
ENSG00000197993	KEL	Kell blood group, metallo-endopeptidase [Source:HGNC Symbol;Acc:6308]	-6.10505537	5.103502598	194.5611546	3.21E-44	8.06E-41	6.075729559	6.416350763	-0.907612816	-0.874832399	4.138027348
ENSG00000143627	PKLR	pyruvate kinase, liver and RBC [Source:HGNC Symbol;Acc:9020]	-7.880567753	4.966265406	188.39622	7.12E-43	1.56E-39	6.184539705	6.111783523	-2.845623393	-7.002070102	3.799873749
ENSG00000170180	GYPA	glycophorin A (MNS blood group) [Source:HGNC Symbol;Acc:4702]	-7.67602921	5.312758052	187.8619216	9.31E-43	1.82E-39	6.067486521	6.812344661	-3.766908349	-2.97427061	4.224973246
ENSG00000197561	ELANE	elastase, neutrophil expressed [Source:HGNC Symbol;Acc:3309]	-6.023520652	5.626150861	183.3194363	9.13E-42	1.60E-38	5.682950487	6.155375067	-1.57307192	-0.406691783	6.942114908
ENSG00000055118	KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2 [Source:HGNC Symbol;Acc:6251]	-5.924295794	6.923850092	181.6691407	2.09E-41	3.34E-38	8.16572184	8.025555626	2.330744516	1.508271876	5.67656232
ENSG00000131747	TOP2A	topoisomerase (DNA) II alpha 170kDa [Source:HGNC Symbol;Acc:11989]	-5.305849613	6.626103217	178.9653596	8.15E-41	1.19E-37	7.516881607	7.660910295	2.1289755	1.926103707	6.641917567
ENSG00000100448	CTSG	cathepsin G [Source:HGNC Symbol;Acc:2532]	-6.007947553	4.87429023	164.0906188	1.45E-37	1.95E-34	5.426778226	5.465512461	-2.845623393	-0.874832399	5.870464764
ENSG00000166947	EPB42	erythrocyte membrane protein band 4.2 [Source:HGNC Symbol;Acc:3381]	-5.874982473	5.431230207	160.912567	7.15E-37	8.96E-34	6.346024108	6.732915862	0.832340008	-1.449512048	4.673975502
ENSG00000211951	IGHV2-26	immunoglobulin heavy variable 2-26 [Source:HGNC Symbol;Acc:5575]	-5.975286828	4.456068793	156.1900471	7.69E-36	9.00E-33	5.80861299	5.379702666	-0.740035668	-1.449512048	3.430160334
ENSG00000211950	IGHV1-24	immunoglobulin heavy variable 1-24 [Source:HGNC Symbol;Acc:5551]	-6.106323229	5.519851109	154.179973	2.12E-35	2.32E-32	7.02191272	6.079121061	0.401591237	-0.589083246	4.886100463
ENSG00000126787	DLGAP5	discs, large (Drosophila) homolog-associated protein 5 [Source:HGNC Symbol;Acc:16864]	-5.030396722	4.916172395	151.7018599	7.36E-35	7.60E-32	6.028383742	5.949930118	0.365283811	0.114810932	4.405023371
ENSG00000039068	CDH1	cadherin 1, type 1, E-cadherin (epithelial) [Source:HGNC Symbol;Acc:1748]	-5.211476345	5.007048086	151.170497	9.62E-35	9.38E-32	6.306294485	6.123695506	0.036330944	0.642123966	3.168221511
ENSG00000243290	IGKV1-12	immunoglobulin kappa variable 1-12 [Source:HGNC Symbol;Acc:5730]	-6.1006739	5.961036193	149.0565118	2.79E-34	2.58E-31	7.514104155	6.586960424	1.098750223	-0.099242291	4.866565467
ENSG00000105610	KLF1	Kruppel-like factor 1 (erythroid) [Source:HGNC Symbol;Acc:6345]	-5.893335408	5.483139827	148.7154177	3.31E-34	2.90E-31	6.696714221	6.563538635	-0.66302287	0.940331404	4.488333455
ENSG00000168754	FAM178B	family with sequence similarity 178, member B [Source:HGNC Symbol;Acc:28036]	-5.635000879	4.527501921	146.0502752	1.27E-33	1.06E-30	5.687439991	5.669431044	-0.99931534	-1.854339469	3.557451401
ENSG00000088325	TPX2	TPX2, microtubule-associated [Source:HGNC Symbol;Acc:1249]	-4.530458654	5.615112454	144.8830773	2.28E-33	1.82E-30	6.672460965	6.601088773	1.89432737	1.820287937	5.263777731
ENSG00000231007	CDC20P1	cell division cycle 20 pseudogene 1 [Source:HGNC Symbol;Acc:29487]	-4.914837736	4.76963257	143.8082952	3.91E-33	2.99E-30	5.992551784	5.715811808	0.663204908	0.466186565	4.077008772
ENSG00000100336	APOL4	apolipoprotein L, 4 [Source:HGNC Symbol;Acc:14867]	-5.404204325	4.551292878	141.9318706	1.01E-32	7.36E-30	5.907365734	5.659645999	-0.059127673	-0.956081001	2.276968566
ENSG00000196188	CTSE	cathepsin E [Source:HGNC Symbol;Acc:2530]	-7.240018267	4.396196948	141.7465507	1.10E-32	7.76E-30	5.275162165	5.882824592	-4.621297774	-3.360122781	2.814341322
ENSG00000206177	HBM	hemoglobin, mu [Source:HGNC Symbol;Acc:4826]	-6.018993929	4.307945359	139.7225011	3.06E-32	2.07E-29	5.402953745	5.522410173	-1.886652327	-3.360122781	3.32911422
ENSG00000075340	ADD2	adducin 2 (beta) [Source:HGNC Symbol;Acc:244]	-4.937267494	6.496103826	138.9965122	4.41E-32	2.87E-29	7.586318744	7.695839822	3.149202463	2.116960417	5.257541658
ENSG00000152078	TMEM56	transmembrane protein 56 [Source:HGNC Symbol;Acc:26477]	-5.101636196	4.71092014	138.7745449	4.93E-32	3.09E-29	5.604410474	5.996020807	-0.907612816	0.192266291	4.039700715
ENSG00000117724	CENPF	centromere protein F, 350/400kDa [Source:HGNC Symbol;Acc:1857]	-4.399528673	5.736076703	137.1273625	1.13E-31	6.85E-29	6.447442448	6.927587689	2.439516168	1.65228493	5.60272092
ENSG00000211937	IGHV2-5	immunoglobulin heavy variable 2-5 [Source:HGNC Symbol;Acc:5576]	-5.788759437	5.939255855	134.1804102	4.99E-31	2.92E-28	7.429800935	6.439885712	1.32354911	-0.053815412	5.453583221

ENSG00000272916	RP11-574K11.31	Bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 2 [Source:UniProtKB/TrEMBL;Acc:S4R438]	1.895169426	-0.080107053	7.425683559	0.006429939	0.045903858	-1.921450438	0.376157587	0.28980717	0.229488395	-1.083770229
ENSG00000102362	SYTL4	synaptotagmin-like 4 [Source:HGNC Symbol;Acc:15588]	-1.508716051	2.529138462	7.42029063	0.00644924	0.046008476	3.279606408	3.624712133	-0.271377577	1.926103707	0.872340852
ENSG00000166682	TMPRSS5	transmembrane protease, serine 5 [Source:HGNC Symbol;Acc:14908]	1.453346791	0.026648253	7.420124482	0.006449835	0.046008476	-0.306880589	0.207316148	-0.059127673	0.154058352	-0.376928291
ENSG00000198682	PAPSS2	3'-phosphoadenosine 5'-phosphosulfate synthase 2 [Source:HGNC Symbol;Acc:8604]	1.187849402	2.942744312	7.418345016	0.006456217	0.046016603	2.093195522	2.865836059	3.186096972	3.462952422	2.693890668
ENSG00000095015	MAP3K1	mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:6848]	1.006237645	7.413284549	7.416767579	0.00646188	0.046019279	6.645948858	7.029244774	7.678502925	8.132325486	7.085254968
ENSG00000180871	CXCR2	chemokine (C-X-C motif) receptor 2 [Source:HGNC Symbol;Acc:6027]	1.153748193	4.447115437	7.417026106	0.006460952	0.046019279	3.774188075	3.594215211	3.963622948	5.031513688	5.118566908
ENSG00000165731	RET	ret proto-oncogene [Source:HGNC Symbol;Acc:9967]	1.429635349	0.540780314	7.416050236	0.006464457	0.046019279	0.254194054	0.583576386	0.036330944	0.798919003	0.615745481
ENSG00000156042	TTC18	tetratricopeptide repeat domain 18 [Source:HGNC Symbol;Acc:30726]	1.303891817	1.387166748	7.404556889	0.006505887	0.046276678	1.063037713	1.202433699	2.06249931	1.538247531	0.415212929
ENSG00000027001	MIPEP	mitochondrial intermediate peptidase [Source:HGNC Symbol;Acc:7104]	-1.20366588	2.461408994	7.402380781	0.006513761	0.046313924	3.265262705	3.100920657	1.907011864	1.006157955	1.71760268
ENSG00000254275	RP11-89M16.1		1.824245891	0.64189445	7.397948715	0.00652983	0.046409376	-0.094677427	0.637745977	2.171650019	-0.956081001	-4.610216385
ENSG00000101986	ABCD1	ATP-binding cassette, sub-family D (ALD), member 1 [Source:HGNC Symbol;Acc:61]	1.007791581	5.118248111	7.392562491	0.006549412	0.046529713	4.323547566	4.848976944	5.034357156	5.666843359	5.356557658
ENSG00000088280	ASAP3	ArfGAP with SH3 domain, ankyrin repeat and PH domain 3 [Source:HGNC Symbol;Acc:14987]	1.362288127	0.764587205	7.390456507	0.006557084	0.046565379	0.627538984	0.637745977	0.959559782	1.069111613	0.095393329
ENSG00000168939	SPRY3	sprouty homolog 3 (Drosophila) [Source:HGNC Symbol;Acc:11271]	1.327178976	1.509170667	7.383842246	0.006581242	0.046718037	1.148882041	0.789025519	1.842443831	1.55300499	1.81594051
ENSG00000115271	GCA	grancalcin, EF-hand calcium binding protein [Source:HGNC Symbol;Acc:15990]	1.056674214	6.854864594	7.381563071	0.006589587	0.04675837	5.762549857	6.144892317	6.431546497	7.293451993	7.716050489
ENSG00000224177	LINC00570	long intergenic non-protein coding RNA 570 [Source:HGNC Symbol;Acc:43717]	-2.195408906	-0.221097885	7.380303926	0.006594203	0.046772212	0.597320328	0.836121102	-2.845623393	-7.002070102	-0.894110176
ENSG00000224997	AL049840.1	Uncharacterized protein; cDNA FLJ53535 [Source:UniProtKB/TrEMBL;Acc:B4DK98]	1.791161532	0.079465802	7.379475133	0.006597242	0.046774873	-0.196880952	-0.713501401	-0.99931534	0.984547783	0.04992943
ENSG00000269430	LRRC3DN	LRRC3 downstream neighbor (non-protein coding) [Source:HGNC Symbol;Acc:1270]	1.530725369	-0.012518843	7.373415318	0.00661951	0.046913803	-0.856736938	-0.112176905	-0.010608983	0.192266291	0.095393329
ENSG00000120278	PLEKHG1	pleckstrin homology domain containing, family G (with RhoGef domain) member 1 [Source:HGNC Symbol;Acc:20884]	-1.045837648	3.990736015	7.371890885	0.006625123	0.046934641	4.659599832	4.343766522	3.159840379	3.47074529	3.765128041
ENSG00000152213	ARL11	ADP-ribosylation factor-like 11 [Source:HGNC Symbol;Acc:24046]	1.178833221	4.211010055	7.365252226	0.006649627	0.047088648	3.270059802	3.408048696	3.599523466	4.81596796	4.990560977
ENSG00000273066	RP11-216L13.19		1.337198964	1.199656576	7.364549306	0.006652227	0.047088648	0.848890306	0.925936776	1.432844522	1.679439471	0.676839958
ENSG00000164509	IL31RA	interleukin 31 receptor A [Source:HGNC Symbol;Acc:18969]	1.637830248	1.177203543	7.362174388	0.006661019	0.047131879	0.796654307	0.689955027	0.210162819	2.106994346	1.134350211
ENSG00000179397	C1orf101	chromosome 1 open reading frame 101 [Source:HGNC Symbol;Acc:28491]	1.422560513	0.185249923	7.359130146	0.006672306	0.047192723	-0.046168491	0.056395802	0.28980717	0.114810932	0.139468177
ENSG00000132470	ITGB4	integrin, beta 4 [Source:HGNC Symbol;Acc:6158]	-1.434396202	1.345867818	7.355262591	0.006686674	0.047256267	2.304339895	2.068117593	-0.161352853	-1.231517591	0.973207307
ENSG00000175061	FAM211A-AS1	FAM211A antisense RNA 1 [Source:HGNC Symbol;Acc:28619]	-1.163828968	8.186586087	7.355523034	0.006685706	0.047256267	8.824219999	8.564105195	8.014408596	7.232878216	7.75287062
ENSG00000257702	LBX2-AS1	LBX2 antisense RNA 1 [Source:HGNC Symbol;Acc:25136]	1.358330496	1.919436653	7.35042267	0.006704699	0.047364585	1.531025871	1.050973916	1.89432737	2.491046607	2.120832889
ENSG00000261996	CTC-281F24.1		1.495877221	0.41714269	7.347814364	0.006714434	0.047395208	-0.489407866	0.498306264	1.120696904	0.722650538	-1.18885156
ENSG00000143061	IGSF3	immunoglobulin superfamily, member 3 [Source:HGNC Symbol;Acc:5950]	-1.401462209	1.004725071	7.344976241	0.006725042	0.04745101	1.883815937	1.997369264	-0.99931534	-0.406691783	-0.376928291
ENSG00000204632	HLA-G	major histocompatibility complex, class I, G [Source:HGNC Symbol;Acc:4964]	1.051900528	4.502240137	7.328995724	0.006785095	0.047836285	3.998873122	3.783314832	4.305306726	5.190971206	4.754122018
ENSG00000105948	TTC26	tetratricopeptide repeat domain 26 [Source:HGNC Symbol;Acc:21882]	-1.330606475	1.839178966	7.32651943	0.00679445	0.047883007	2.798442909	2.369639072	0.6928072	0.265774258	1.410986673
ENSG00000117425	PTCH2	patched 2 [Source:HGNC Symbol;Acc:9586]	1.383549148	0.448712955	7.302763092	0.006884865	0.048500727	0.254194054	0.468724906	0.6928072	0.434658742	0.002986071
ENSG00000169258	GPRIN1	G protein regulated inducer of neurite outgrowth 1 [Source:HGNC Symbol;Acc:24835]	1.400363076	0.551438443	7.301219368	0.006890783	0.048522943	0.090278201	0.498306264	0.505333341	0.669468625	0.646616095
ENSG00000177096	FAM109B	family with sequence similarity 109, member B [Source:HGNC Symbol;Acc:27161]	-1.124077846	3.066512988	7.297885336	0.006903581	0.048592364	3.701474281	3.541909135	1.956663165	2.404187718	2.956635145
ENSG00000157657	ZNF618	zinc finger protein 618 [Source:HGNC Symbol;Acc:29416]	-1.094603082	3.373908118	7.297209782	0.006906178	0.048592364	3.878388807	3.847931966	2.213098418	2.756300061	3.465217709